Condition estimation of linear algebraic equations and its application to feature selection

Joab Winkler

Department of Computer Science, The University of Sheffield, Sheffield, United Kingdom

The Institute of High Performance Computing
A* Agency for Science, Technology and Research
Singapore

January 2019

1. Introduction
   - Mathematical background
2. Regression
3. Condition numbers and regularisation
   - The effective condition number
   - The discrete Picard condition
   - Tikhonov regularisation
   - Componentwise condition numbers
4. Feature selection
   - Feature selection and condition estimation
5. Summary

Introduction

Several problems require the prediction of the output of a physical system for which the sample size n is much smaller than the dimension of the data p:

- Chemometrics
- Brain imaging
- Genomics
- Gene selection from microarray data
- Text analysis

The condition $n < p$ implies that there are many models that satisfy the given data and important issues therefore arise:

- Which model from this infinite set of models should be chosen?
- What is the criterion that should be used for this selection?
- Can the selection be generic, that is, not problem dependent, such that prior information is not required?

Mathematical background

These problems yield an equation of the form

$$Ax = b + \varepsilon, \qquad A \in \mathbb{R}^{m\times n}, \qquad b \in \mathbb{R}^m, \qquad x \in \mathbb{R}^n$$

where $m < n$, $\operatorname{rank} A = m$ and ε is the noise. The least squares minimisation of $\|\varepsilon\|$ leads to the normal equation

$$A^T A x = A^T b$$

whose solution is

$$x_{\mathrm{soln}} = A^\dagger b = V S^\dagger U^T b$$

where the superscript $\dagger$ denotes the pseudo-inverse. The solution is

$$x_{\mathrm{soln}} = x_{\mathrm{ln}} + x_0, \quad x_{\mathrm{ln}} = V \begin{bmatrix} S_1^{-1} U^T b \\ 0_{n-m} \end{bmatrix}, \quad x_0 = V \begin{bmatrix} 0_m \\ r \end{bmatrix}$$

where x_{ln} is the minimum norm solution, x_0 lies in the null space of A, r is arbitrary, and

$$S = \begin{bmatrix} S_1 & 0_{m,n-m} \end{bmatrix}$$

The solution x_{ln} is unsatisfactory for two reasons:

- *Prediction accuracy*: This solution may have low bias and high variance. Prediction accuracy can sometimes be improved by reducing or setting to zero some coefficients of x_{ln}.
- *Interpretation*: It is usually desirable to choose the most important components of x_{ln} that characterise the physical system being considered.

Methods that are used to overcome these problems:

- *Ridge regression*: The magnitude of the components of x_{ln} is reduced continuously:
  - It is more stable than subset selection.
  - It does not set any components to zero and thus it does not yield a sparse model that can be easily interpreted.
- *Subset selection*: Components of x_{ln} are deleted in discrete steps:
  - The models are strongly dependent on the components that are deleted because the elimination procedure is discrete.
  - A small change in the data can cause a large change in the selected model, which reduces the prediction accuracy.

Ridge regression (Tikhonov regularisation)

The sensitivity of the solution x_{ln} to perturbations in b can be reduced by a constraint on the magnitude of the solution x_{reg}:

$$x_{\mathrm{reg}} = \arg\min_{x} \left\{ (Ax - b)^T (Ax - b) + \lambda \|x\|^2 \right\}, \qquad \lambda > 0$$

and thus

$$\left(A^T A + \lambda I\right) x_{\mathrm{reg}} = A^T b, \qquad \lambda > 0$$

The lasso ('least absolute shrinkage and selection operator')

The retension of the advantages of ridge regression (stability) and subset selection (sparsity) are combined in the lasso:

$$x_{\mathrm{lasso}} = \arg\min_{x} \left\{ (Ax - b)^T (Ax - b) \right\} \quad \text{subject to} \quad \|x\|_1 \leq t$$

which can also be written as

$$x_{\mathrm{lasso}} = \arg\min_{x} \left\{ (Ax - b)^T (Ax - b) + \lambda \|x\|_1 \right\}, \qquad \lambda > 0$$

The elastic net

This method is an improvement on the lasso and it combines L_1 and L_2 regularisation:

$$x_{\mathrm{elastic}} = \arg\min_x \left\{ (Ax - b)^T (Ax - b) + \lambda_1 \|x\|_1 + \lambda_2 \|x\|^2 \right\}$$

where

$$\lambda_1, \lambda_2 > 0$$

The solutions from Tikhonov regularisation, the lasso and the elastic net reduce the sensitivity of the least norm solution x_{ln} to perturbations in b, but there are differences between these forms of regularisation.

Compare Tikhonov regularisation

- Tikhonov regularisation imposes a Gaussian prior on the parameters of the model.
- Tikhonov regularisation does not impose sparsity on x_{reg}.
- The solution x_{reg} has a closed form expression.

with the lasso

- The lasso imposes a Laplacian prior on the parameters of the model.
- The lasso favours sparse solutions because some coefficients of x_{lasso} are set to zero. The sparsity of x_{lasso} increases as λ increases.
- The solution x_{lasso} does not have a closed form expression and quadratic programming is required for its computation.

and the elastic net

- The sparsity of x_{elastic} is similar to the sparsity of x_{lasso}.
- The solution x_{elastic} favours a model in which strongly correlated predictors are usually either all included, or all excluded.
- The solution x_{elastic} is much better than x_{lasso} for some problems.

A regularised solution (Tikhonov, the lasso and the elastic net) is stable with respect to perturbations in b, but several points arise:

- Is regularisation *always* required when the data b are corrupted by noise?
- Must specific conditions on A and b be satisfied in order that regularisation is imposed only when it is required?
- What are consequences of applying regularisation when it is not required?
- If regularisation is required, then

  $$r_{\mathrm{method}} = x_{\mathrm{ln}} - x_{\mathrm{method}} \neq 0, \quad \mathrm{method} = \{\mathrm{reg},\ \mathrm{lasso},\ \mathrm{elastic}\}$$

  Can bounds be imposed on $\|r_{\mathrm{reg}}\|$, $\|r_{\mathrm{lasso}}\|$ and $\|r_{\mathrm{elastic}}\|$, such that these errors induced by regularisation are quantified?

The answers to these questions are most easily obtained if Tikhonov regularisation is considered because the constraint in the 2-norm lends itself naturally to the SVD.

Regression

The use of regularisation is usually justified for three reasons:

- It reduces or eliminates over-fitting in regression.
- It reduces the sensitivity of the regression curve to noise in the data.
- It imposes a unique solution in feature selection

$$Ax = b$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $x \in \mathbb{R}^n$, $m < n$ and $\operatorname{rank} A = m$.

But

- There are well-defined problems for which regularisation must not be used because it causes a large degradation in the solution.

and thus

- Can a quantitative test be established such that regularisation is used only when it is required?

Regression provides a good example of the correct use, and the incorrect use, of regularisation.

Example 1 Consider the points (x_i, y_i), $i = 1, \dots, 100$, where the independent variables x_i are not uniformly distributed in the interval $I = [1, \dots, 20]$, the dependent variables y_i are given by

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

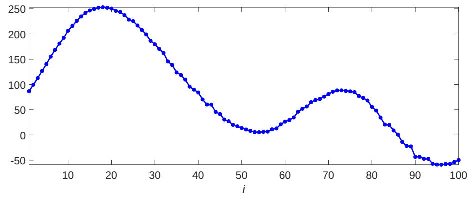
the centres d_k of the 33 basis functions are uniformly distributed in I and $\sigma_d = 1.35$.

Consider two sets of data points, $y = y_1$ and $y = y_2$, and the perturbations δy_1 and δy_2,

$$\delta y_1, \delta y_2 \sim \mathcal{N}\left(\mu = 0, \sigma^2 = 25 \times 10^{-8}\right)$$

and

$$\frac{\|\delta y_1\|}{\|y_1\|} = 3.41 \times 10^{-6} \qquad \text{and} \qquad \frac{\|\delta y_2\|}{\|y_2\|} = 8.27 \times 10^{-4}$$



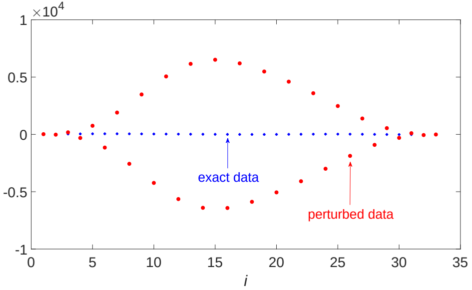


Figure: The exact curve (top), and the coefficients a_i (bottom) for the exact data $y = y_1$ and the perturbed data $y = y_1 + \delta y_1$.

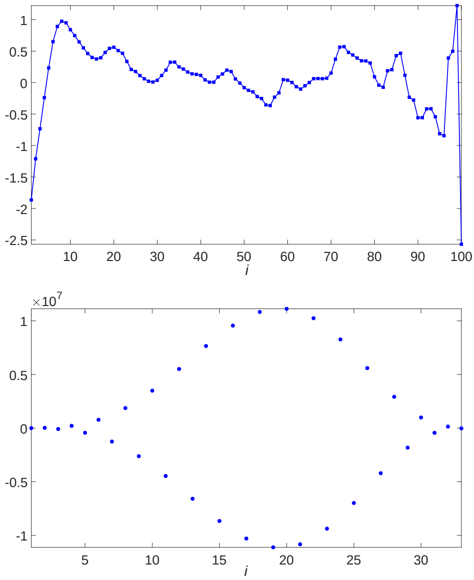


Figure: The exact curve (top), and the coefficients $\log_{10}|a_i|$ (bottom) for the exact data $y = y_2$ and the perturbed data $y = y_2 + \delta y_2$.

- The interpolated curve is unstable for the data set $y = y_1$.
- The interpolated curve is stable for the data set $y = y_2$.

The coefficient matrix $A \in \mathbb{R}^{100\times 33}$ is the same for $y = y_1$ and $y = y_2$, and its condition number is

$$\kappa(A) = 5.11 \times 10^8$$

Thus

- *The presence of noise in the vector b, where $Ax = b$, does not imply that x is sensitive to changes in b.*
- *The condition $\kappa(A) \gg 1$ does not imply that the equation $Ax = b$ is ill-conditioned.*
- Tikhonov regularisation yields a very good result (numerically stable and a small error between the theoretically exact solution and the regularised solution) for $y = y_1$, but an unsatisfactory result for $y = y_2$ (a very large error between the theoretically exact solution and the regularised solution).

Condition numbers and regularisation

The 2-norm condition number of $A \in \mathbb{R}^{m \times n}$ is

$$\kappa(A) = \frac{s_1}{s_p}, \qquad p = \min(m, n)$$

where $s_i, i = 1, \ldots, p$, are the singular values of A and $\operatorname{rank} A = p$.

- The condition number $\kappa(A)$ cannot be a measure of the stability of $Ax = b$ because it is independent of b.
- It is necessary to develop a measure of stability that is a function of A and b.
- This leads to:
  - A refined normwise condition number - the effective condition number - which is a function of A and b.
  - Componentwise condition numbers - one condition number for each component of x.

The effective condition number

The effective condition number $\eta(A, b)$ of

$$A^T Ax = A^T b, \qquad A \in \mathbb{R}^{m\times n}, \qquad m \geq n$$

is a refined normwise condition number.

Theorem 1 Let the relative errors Δx and Δb be

$$\Delta x = \frac{\|\delta x\|}{\|x\|} \qquad \text{and} \qquad \Delta b = \frac{\|\delta b\|}{\|b\|}$$

The effective condition number $\eta(A, b)$ of $A^T Ax = A^T b$ is equal to the maximum value of the ratio of Δx to Δb with respect to all perturbations $\delta b \in \mathbb{R}^m$

$$\eta(A, b) = \max_{\delta b \in \mathbb{R}^m} \frac{\Delta x}{\Delta b} = \frac{1}{s_n} \frac{\|c\|}{\|S^\dagger c\|}$$

where $A = USV^T$ is the SVD of A and $c = U^T b$.

Proof It follows from

$$A^T Ax = A^T b$$

that $x = VS^\dagger U^T b = VS^\dagger c$ and thus

$$\|\delta x\| = \left\|VS^\dagger U^T \delta b\right\| = \left\|S^\dagger \delta c\right\| \leq \frac{\|\delta c\|}{s_n}$$

and the division of both sides of this inequality by $\|x\|$ yields

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{1}{s_n}\frac{\|\delta c\|}{\|c\|}\frac{\|c\|}{\|x\|} = \frac{1}{s_n}\frac{\|\delta b\|}{\|b\|}\frac{\|c\|}{\|x\|}$$

and thus

$$\eta(A, b) = \max_{\delta b \in \mathbb{R}^m} \frac{\Delta x}{\Delta b} = \frac{1}{s_n}\frac{\|c\|}{\|S^\dagger c\|} = \max_{\delta b \in \mathbb{R}^m} \frac{\Delta x}{\Delta b} = \frac{1}{s_n}\frac{\|b\|}{\|S^\dagger U^T b\|}$$

□

$$\eta(A,b) = \max_{\delta b \in \mathbb{R}^m} \frac{\Delta x}{\Delta b} = \frac{1}{s_n} \frac{\|c\|}{\|S^\dagger c\|} = \max_{\delta b \in \mathbb{R}^m} \frac{\Delta x}{\Delta b} = \frac{1}{s_n} \frac{\|b\|}{\|S^\dagger U^T b\|}$$

- The minimum value of $\eta(A,b)$ occurs when $c = e_n$ (e_i is the ith unit basis vector).
- If some conditions on b are satisfied, the maximum value of $\eta(A,b)$ occurs when $c = e_1$, and thus

$$1 \leq \eta(A,b) \leq \frac{s_1}{s_n} = \kappa(A)$$

More generally, since $b = Uc$:

- $\eta(A,b) \approx 1$ if the dominant components of b lie along the columns u_i of U that are defined by large values of i.
- $\eta(A,b) \approx \kappa(A)$ if the dominant components of b lie along the columns u_i of U that are defined by small values of i.

The discrete Picard condition

$$\eta(A, b) = \max_{\delta b \in \mathbb{R}^m} \frac{\Delta x}{\Delta b} = \frac{1}{s_n} \frac{\|c\|}{\|S^\dagger c\|}, \qquad c = U^T b$$

Consider the ratio $|c_i|/s_i$, $i = 1, \ldots, n$.

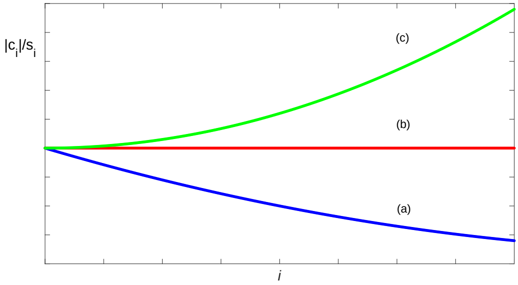


Figure: Three forms for the ratio $|c_i|/s_i$: (a) monotonically decreasing, (b) constant and (c) monotonically increasing.

- If ${|c_i|}/{s_i} \to 0$ as $i \to n$, then the constants $|c_i|$ decay to zero faster than the singular values s_i. The condition

$$\frac{|c_i|}{s_i} \to 0 \qquad \text{as} \qquad i \to n$$

is called the discrete Picard condition. If this condition is satisfied, then $\eta(A, b) \approx \kappa(A)$ and thus $Ax = b$ is ill-conditioned. Tikhonov regularisation can be used to yield a well-conditioned solution x.

- If ${|c_i|}/{s_i} \approx 1$, $i = 1, \ldots, n$, then $\eta(A, b) \approx {\kappa(A)}/{\sqrt{n}}$. Tikhonov regularisation cannot be used to yield a well-conditioned solution x.
- If

$$|c_{i+1}| \gg |c_i|, \qquad i = 1, \ldots, n-1$$

then $\eta(A, b) \approx 1$ and regularisation must not be applied.

Computation of the discrete Picard condition

The important term for the evaluation of the discrete Picard condition is

$$\|x\| = \left\|A^\dagger b\right\| = \left\|S^\dagger c\right\|$$

If noise is present, then the square of the magnitude of the perturbed solution is

$$\|x + \delta x\|^2 = \left\|A^\dagger (b + \delta b)\right\|^2 = \left\|S^\dagger (c + \delta c)\right\|^2 = \sum_{i=1}^{n} \left(\frac{c_i + \delta c_i}{s_i}\right)^2$$

Assume the magnitude of the perturbation $|\delta c_i|$ is approximately constant, such that

$$\begin{array}{lcll} |\delta c_i| & \ll & |c_i|\,, & i = 1, \dots, p-1 \\ |\delta c_i| & \approx & |c_p|\,, & i = p \\ |\delta c_i| & \gg & |c_i|\,, & i = p+1, \dots, n \end{array}$$

It follows that if the discrete Picard condition is satisfied

$$\frac{|c_i + \delta c_i|}{s_i} \approx \begin{cases} \frac{|c_i|}{s_i}, & i = 1, \ldots, p-1 \\ \frac{|c_p + s_p|}{s_p}, & i = p \\ \frac{|\delta c_i|}{s_i}, & i = p+1, \ldots, n \end{cases}$$

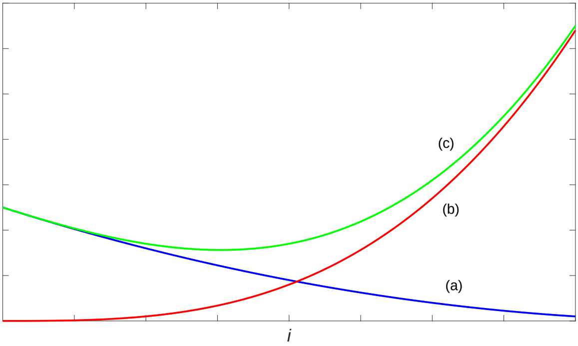


Figure: The ratios (a) $|c_i|/s_i$, (b) $|\delta c_i|/s_i$ and (c) $|c_i+\delta c_i|/s_i$ if the discrete Picard condition is satisfied.

- If the discrete Picard condition

$$\frac{|c_i|}{s_i} \to 0 \qquad \text{as} \qquad i \to n$$

is satisfied, the function ${}^{|c_i+\delta c_i|}/_{s_i}$ cannot be computed because it is sensitive to noise.

- A similar result occurs if

$$|c_i| \approx s_i, \qquad i = 1, \ldots, n$$

because

$$\frac{|c_i + \delta c_i|}{s_i} \approx \frac{|\delta c_i|}{s_i}, \qquad i = p+1, \ldots, n$$

- If

$$|c_{i+1}| \gg |c_i|, \qquad i = 1, \ldots, n-1$$

computational problems do not occur because

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

Summary (discrete Picard condition)

- The form of the term $|c_i|/s_i$ defines the stability of $Ax = b$.
- If

$$\frac{|c_i|}{s_i} \to 0 \qquad \text{as} \qquad i \to n$$

  then $Ax = b$ is ill-conditioned. The dominant components of b lie along the columns u_i of U that are defined by small values of i (large singular values). Tikhonov regularisation enables a well-conditioned solution to be computed.
- If

$$|c_{i+1}| \gg |c_i|, \quad i = 1, \ldots, n-1$$

  then $Ax = b$ is well-conditioned and $|c_i|/s_i$ is numerically stable. The dominant components of b lie along the columns u_i of U that are defined by large values of i (small singular values).

Example 2 Consider the regression problem in Example 1.

- The data points $y = y_1$ yield an ill-conditioned equation $A^T Ax = A^T b$. The effective condition number is $\eta(A, b) = 4.61 \times 10^8$.
- The data points $y = y_2$ yield a well-conditioned equation $A^T Ax = A^T b$. The effective condition number is $\eta(A, b) = 7.94$.

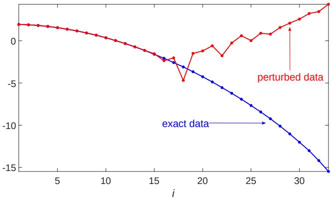


Figure: The ratio $\log_{10} {|c_i|}/{s_i}$ for the exact data $y = y_1$ and the ratio $\log_{10} {|c_i + \delta c_i|}/{s_i}$ for the perturbed data $y = y_1 + \delta y_1$.

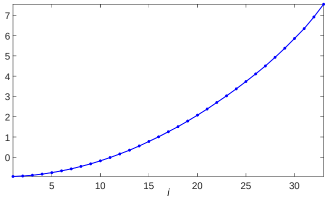


Figure: The ratio $\log_{10} {|c_i|}/{s_i}$ for the exact data $y = y_2$ and the ratio $\log_{10} {|c_i + \delta c_i|}/{s_i}$ for the perturbed data $y = y_2 + \delta y_2$.

Tikhonov regularisation

If the discrete Picard condition is satisfied, the equation $A^TAx = A^Tb$ is ill-conditioned and Tikhonov regularisation requires the solution $x(\lambda)$ of

$$(A^TA + \lambda I)x(\lambda) = A^Tb$$

and thus

$$x(\lambda) = V\left((S^TS + \lambda I)^{-1}S^T\right)U^Tb, \qquad \lambda \geq 0$$

- What is the error between the regularised solution $(\lambda > 0)$ and the exact solution $(\lambda = 0)$ if:
  - the discrete Picard condition is satisfied
  - the discrete Picard condition is not satisfied

Assume that λ^* is the optimal value of the regularisation parameter λ.

- If the discrete Picard condition, $\frac{|c_i|}{s_i} \to 0$ as $i \to n$, is satisfied, the solution $x(0)$ is dominated by the large singular values and it is independent of the small singular values. The error is small

$$\frac{\|x(\lambda^*) - x(0)\|}{\|x(0)\|} \approx \frac{\lambda^*}{\lambda^* + s_1^2} \ll 1$$

because Tikhonov regularisation filters out the small singular values.

- If $|c_i| \approx s_i, i = 1, \ldots, n$, the error is smaller because all the singular values contribute to $x(0)$

$$\frac{\|x(\lambda^*) - x(0)\|}{\|x(0)\|} \approx \left(\frac{n-p}{n}\right)^{\frac{1}{2}}$$

Tikhonov regularisation filters out the components of $x(0)$ associated with the small singular values, but the components associated with the large singular values are not affected.

- If ${}^{|c_{i+1}|}/_{s_{i+1}} \gg {}^{|c_i|}/_{s_i}, i = 1, \ldots, n-1$, the error is large

$$\frac{\|x(\lambda^*) - x(0)\|}{\|x(0)\|} \approx 1, \qquad \|x(\lambda^*)\| \approx 0$$

because $x(0)$ is dominated by the small singular values, all of which are filtered out by Tikhonov regularisation.

Componentwise condition numbers

The condition number $\kappa(A)$ and effective condition number $\eta(A, b)$ are defined in the normwise sense. Condition numbers defined in the componentwise sense are more refined.

- A condition number is assigned to *each component* of x, where $A^T Ax = A^T b$.

The condition number $\rho(A, b, x_j)$ of the component x_j of x is defined as

$$\rho(A, b, x_j) = \max_{\delta b \in \mathbb{R}^m} \frac{\Delta x_j}{\Delta b}, \qquad j = 1, \ldots, n$$

where

$$\Delta x_j = \frac{|\delta x_j|}{|x_j|} \qquad \text{and} \qquad \Delta b = \frac{\|\delta b\|}{\|b\|}$$

Theorem 2 The condition number $\rho(A, b, x_j)$ of the component x_j, $j = 1, \ldots, n$, of $A^T Ax = A^T b$ is

$$\rho(A, b, x_j) = \frac{\left\| e_j^T A^\dagger \right\| \|b\|}{|x_j|} = \frac{\left\| e_j^T A^\dagger \right\| \|b\|}{\left| e_j^T A^\dagger b \right|} = \frac{1}{\cos \gamma_j} > 1$$

where γ_j is the angle between b and the jth row of $A^\dagger$.

Proof If e_j is the jth unit basis vector, then a change δb in b causes a change δx_j in x_j

$$\delta x_j = e_j^T \delta x = e_j^T A^\dagger \delta b, \qquad j = 1, \ldots, n$$

and thus

$$\Delta x_j = \frac{|\delta x_j|}{|x_j|} \leq \frac{\left\| e_j^T A^\dagger \right\| \|\delta b\|}{|x_j|}, \qquad j = 1, \ldots, n$$

and the result follows. $\square$

Example 3 Consider the regression problem in Example 1.

- The data points $y = y_1$ yield an ill-conditioned equation $A^T Ax = A^T b$.
- The data points $y = y_2$ yield a well-conditioned equation $A^T Ax = A^T b$.

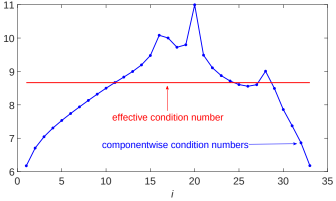


Figure: The effective condition number and the componentwise condition numbers, on a logarithmic scale, for $y = y_1$.

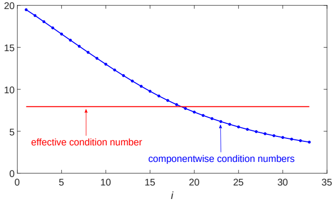


Figure: The effective condition number and the componentwise condition numbers for $y = y_2$.

Feature selection

Many problems in feature selection yield the equation

$$Ax = b + \varepsilon$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $x \in \mathbb{R}^n$, $m < n$, $\operatorname{rank} A = m$ and ε is the noise in the data b. The normal equations are

$$\left(A^T A\right) x = A^T b$$

This equation has an infinite number of solutions x_{soln}

$$x_{\mathrm{soln}} = V \begin{bmatrix} S_1^{-1} U^T b \\ 0_{n-m} \end{bmatrix} + V \begin{bmatrix} 0_m \\ r \end{bmatrix} = x_{\mathrm{ln}} + x_0$$

where x_{ln} is the solution of minimum norm, x_0 lies in the null space of A and is orthogonal to x_{ln}, $r \in \mathbb{R}^{n-m}$ is arbitrary and

$$A = USV^T = U \begin{bmatrix} S_1 & 0_{m,n-m} \end{bmatrix} V^T, \qquad S_1 = \mathrm{diag}\left\{s_i\right\}_{i=1}^m$$

- The most important features (components of x_{soln}) of the system are usually sought and a sparse solution is therefore desired.
- Use the $n-m$ degrees of freedom in r in the null space vector x_0 to obtain a sparse and regularised solution x_{soln}

$$x_{\mathrm{soln}} = V \begin{bmatrix} S_1^{-1} U^T b \\ 0_{n-m} \end{bmatrix} + V \begin{bmatrix} 0_m \\ r \end{bmatrix} = x_{\mathrm{ln}} + x_0$$

  The solutions x_{lasso} and x_{elastic} ignore the vector x_0.
- The vector r is chosen such that if the d components $k_1, k_2, \ldots, k_d$, of x_{ln} have large condition numbers, then these d components of the solution x_{soln} are equal to zero.
- The lasso and the elastic net yield sparse solutions from constrained minimisation problems but they are *approximate* solutions of the normal equations

$$\left(A^T A\right) x = A^T b$$

  The need, or otherwise, for regularisation, and the errors between (a) the solutions x_{soln} and x_{lasso}, and (b) the solutions x_{soln} and x_{elastic}, are not considered.

Feature selection and condition estimation

Since

$$x_{\mathrm{soln}} = x_{\mathrm{ln}} + x_0 = x_{\mathrm{ln}} + V \begin{bmatrix} 0_m \\ r \end{bmatrix} = x_{\mathrm{ln}} + \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} 0_m \\ r \end{bmatrix}$$

it follows that

$$x_{\mathrm{soln}} = x_{\mathrm{ln}} + V_2 r, \qquad V_2 \in \mathbb{R}^{n \times (n-m)}$$

By definition, the components of x_{ln} that have the d largest condition numbers, $1 \leq d \leq n - m$, satisfy

$$x_{\mathrm{ln},k} + x_{0,k} = 0, \qquad k = k_1, k_2, \ldots, k_d$$

that is, these components of x_{soln} are set to zero, thereby inducing sparsity in x_{soln}.

- The vector r is chosen to satisfy these d equations.

Summary

- A large condition number of A, $\kappa(A) \gg 1$, does not imply that the equation $Ax = b$ is ill-conditioned.
- The effective condition number $\eta(A, b)$ is a better measure of the stability of $Ax = b$ because it is a function of A and b, which must be compared with $\kappa(A)$, which is a function of A only.
- If the discrete Picard condition is satisfied, then $\eta(A, b) \approx \kappa(A)$ and Tikhonov regularisation yields a stable solution whose error is small.
- More refined measures of the stability of $Ax = b$ are the componentwise condition numbers, which measure the stability of each component of x due to a perturbation in b.
- The equation $Ax = b$ for feature selection has an infinite number of solutions. The lasso and elastic net apply regularisation to the least norm solution x_{ln}, and they do not consider the vectors x_{null} that lie in the null space of A.
- Consideration of the componentwise condition numbers of $x_{\mathrm{soln}} = x_{\mathrm{ln}} + x_{\mathrm{null}}$ yields a sparse solution that satisfies the normal equations, and regularisation is satisfied in the componentwise sense.