

# Statistical Inverse Problems

## Example 1: Regression (R. S. Wilson)

Axel Munk

January 2010

## Linear regression model

$$Y = Ax + \epsilon \quad Y = (Y_1, \dots, Y_n)^T \in \mathbb{R}^n$$
$$\epsilon \in \mathbb{R}^n \quad \epsilon_j \stackrel{iid}{\sim} \langle 0, \sigma^2 \rangle$$

Assume  $x \in \mathbb{R}^n$ ,  $A \in \mathbb{R}^{n \times n}$  invertible (for simplicity)

## Linear regression model

$$Y = Ax + \epsilon \quad Y = (Y_1, \dots, Y_n)^T \in \mathbb{R}^n$$
$$\epsilon \in \mathbb{R}^n \quad \epsilon_j \stackrel{iid}{\sim} \langle 0, \sigma^2 \rangle$$

Assume  $x \in \mathbb{R}^n$ ,  $A \in \mathbb{R}^{n \times n}$  invertible (for simplicity)

LSE (MLE for  $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$ ,  $\epsilon_j \stackrel{iid}{\sim} N(0, \sigma^2)$ )

$$\hat{x} = \operatorname{argmin}_{x \in \mathbb{R}^n} \|Y - Ax\|^2 \quad (\text{often write } \|\cdot\| \text{ for } \|\cdot\|_n)$$

## Linear regression model

$$Y = Ax + \epsilon \quad Y = (Y_1, \dots, Y_n)^T \in \mathbb{R}^n$$
$$\epsilon \in \mathbb{R}^n \quad \epsilon_j \stackrel{iid}{\sim} \langle 0, \sigma^2 \rangle$$

Assume  $x \in \mathbb{R}^n$ ,  $A \in \mathbb{R}^{n \times n}$  invertible (for simplicity)

LSE (MLE for  $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$ ,  $\epsilon_j \stackrel{iid}{\sim} N(0, \sigma^2)$ )

$$\hat{x} = \operatorname{argmin}_{x \in \mathbb{R}^n} \|Y - Ax\|^2 \quad (\text{often write } \|\cdot\| \text{ for } \|\cdot\|_n)$$

$$\text{gives here: } \hat{x} = (A^T A)^{-1} A^T Y = A^{-1} (A^T)^{-1} A^T Y = A^{-1} Y$$

## Linear regression model

$$Y = Ax + \epsilon \quad Y = (Y_1, \dots, Y_n)^T \in \mathbb{R}^n$$
$$\epsilon \in \mathbb{R}^n \quad \epsilon_j \stackrel{iid}{\sim} \langle 0, \sigma^2 \rangle$$

Assume  $x \in \mathbb{R}^n$ ,  $A \in \mathbb{R}^{n \times n}$  invertible (for simplicity)

$$\text{LSE (MLE for } \epsilon = (\epsilon_1, \dots, \epsilon_n)^T, \epsilon_j \stackrel{iid}{\sim} N(0, \sigma^2))$$

$$\hat{x} = \operatorname{argmin}_{x \in \mathbb{R}^n} \|Y - Ax\|^2 \quad (\text{often write } \|\cdot\| \text{ for } \|\cdot\|_n)$$

$$\text{gives here: } \hat{x} = (A^T A)^{-1} A^T Y = A^{-1} (A^T)^{-1} A^T Y = A^{-1} Y$$

$\sigma^2 = 0$  (error free model)  $\Rightarrow$  compute  $x = A^{-1} Y$

## Linear regression model

$$Y = Ax + \epsilon \quad Y = (Y_1, \dots, Y_n)^T \in \mathbb{R}^n$$
$$\epsilon \in \mathbb{R}^n \quad \epsilon_j \stackrel{iid}{\sim} \langle 0, \sigma^2 \rangle$$

Assume  $x \in \mathbb{R}^n$ ,  $A \in \mathbb{R}^{n \times n}$  invertible (for simplicity)

$$\text{LSE (MLE for } \epsilon = (\epsilon_1, \dots, \epsilon_n)^T, \epsilon_j \stackrel{iid}{\sim} N(0, \sigma^2))$$

$$\hat{x} = \operatorname{argmin}_{x \in \mathbb{R}^n} \|Y - Ax\|^2 \quad (\text{often write } \|\cdot\| \text{ for } \|\cdot\|_n)$$

$$\text{gives here: } \hat{x} = (A^T A)^{-1} A^T Y = A^{-1} (A^T)^{-1} A^T Y = A^{-1} Y$$

$\sigma^2 = 0$  (error free model)  $\Rightarrow$  compute  $x = A^{-1} Y$

n=4:

Let  $A = A^T = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}$   $b = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}$

solution  $x^* = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$

Consider a small deterministic perturbation (measurement error)

- $\epsilon \sim 0.1$

$$Y = \begin{pmatrix} 32 \\ \boxed{22.9} \\ 33 \\ 31 \end{pmatrix} \Rightarrow \hat{x} = \begin{pmatrix} 5.1 \\ -5.8 \\ 2.7 \\ -5 \cdot 10^{-14} \end{pmatrix} \quad (1)$$

- $\epsilon \sim 0.01$

still

$$\hat{x} = \begin{pmatrix} 1.41 \\ 0.32 \\ 1.17 \\ 0.9 \end{pmatrix} \quad (2)$$

Consider a small deterministic perturbation (measurement error)

- $\epsilon \sim 0.1$

$$Y = \begin{pmatrix} 32 \\ \boxed{22.9} \\ 33 \\ 31 \end{pmatrix} \Rightarrow \hat{x} = \begin{pmatrix} 5.1 \\ -5.8 \\ 2.7 \\ -5 \cdot 10^{-14} \end{pmatrix} \quad (1)$$

- $\epsilon \sim 0.01$

still

$$\hat{x} = \begin{pmatrix} 1.41 \\ 0.32 \\ 1.17 \\ 0.9 \end{pmatrix} \quad (2)$$

The error in  $\hat{x}$  is of the order  $10 \cdot \epsilon$ .

Consider a small deterministic perturbation (measurement error)

- $\epsilon \sim 0.1$

$$Y = \begin{pmatrix} 32 \\ \boxed{22.9} \\ 33 \\ 31 \end{pmatrix} \Rightarrow \hat{x} = \begin{pmatrix} 5.1 \\ -5.8 \\ 2.7 \\ -5 \cdot 10^{-14} \end{pmatrix} \quad (1)$$

- $\epsilon \sim 0.01$

still

$$\hat{x} = \begin{pmatrix} 1.41 \\ 0.32 \\ 1.17 \\ 0.9 \end{pmatrix} \quad (2)$$

The error in  $\hat{x}$  is of the order  $10 \cdot \epsilon$ .

$$\text{MSE}_{x_i} [\hat{x}_i] = E |\hat{x}_i - x_i|^2 \quad i = 1, \dots, 4$$

We know

$$E\hat{x} = A^{-1}EY = A^{-1}Ax = x \quad (\text{unbiased})$$

hence

$$\text{MSE} [\hat{x}_i] = \text{Var} [\hat{x}_i],$$

$$\begin{aligned} \text{Cov} [\hat{x}] &= \text{Cov} [A^{-1}Y] = \sigma^2 A^{-1}I (A^{-1})^T \\ &= \sigma^2 (A^T A)^{-1} \end{aligned}$$

$$\begin{array}{c} \text{Var}[\hat{x}_2] \\ \downarrow \\ \text{Cov}[\hat{x}] = \sigma^2 (A^T A)^{-1} = \sigma^2 \begin{pmatrix} 2442 & \dots & & & \\ \dots & \boxed{6694} & \dots & & \\ & \dots & 423 & \dots & \\ & & \dots & 149 & \end{pmatrix} \end{array}$$

- $\sigma = 0.1$   
error of second component  $\approx 0.1 \cdot \sqrt{6694} \approx 8.2$  (cf. (1))
- $\sigma = 0.01$   
error of second component  $\approx 0.82$
- Roughly:  $10 \times \sigma$

Hence, in a sense, the  $\text{Cov}[\hat{x}]$  measures the "loss in precision" when inverting a matrix.

$$\begin{array}{c} \text{Var}[\hat{x}_2] \\ \downarrow \\ \text{Cov}[\hat{x}] = \sigma^2 (A^T A)^{-1} = \sigma^2 \begin{pmatrix} 2442 & \dots & & & \\ \dots & \boxed{6694} & \dots & & \\ & \dots & 423 & \dots & \\ & & \dots & 149 & \end{pmatrix} \end{array}$$

- $\sigma = 0.1$   
error of second component  $\approx 0.1 \cdot \sqrt{6694} \approx 8.2$  (cf. (1))
- $\sigma = 0.01$   
error of second component  $\approx 0.82$
- Roughly:  $10 \times \sigma$

Hence, in a sense, the  $\text{Cov}[\hat{x}]$  measures the "loss in precision" when inverting a matrix.

## How is this done by numerical analysts?

Perturbation theory: (K. Lange, p.75)

$$Ax = b \qquad A(x + \Delta x) = b + \Delta b$$

let  $\|A\|$  the matrix norm (operator norm, spectral norm)

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|$$

$$\begin{aligned} \overset{A \text{ invertible}}{\Rightarrow} \quad \|b\| &\leq \|A\| \cdot \|x\| & (a) \\ \|\Delta x\| &\leq \|A^{-1}\| \cdot \|\Delta b\| & (b) \end{aligned}$$

## How is this done by numerical analysts?

Perturbation theory: (K. Lange, p.75)

$$Ax = b \qquad A(x + \Delta x) = b + \Delta b$$

let  $\|A\|$  the matrix norm (operator norm, spectral norm)

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|$$

$$\begin{aligned} A \text{ invertible} \\ \Rightarrow \quad \|b\| &\leq \|A\| \cdot \|x\| & (a) \\ \| \Delta x \| &\leq \|A^{-1}\| \cdot \| \Delta b \| & (b) \end{aligned}$$

hence

$$\frac{(a)}{(b)} \Leftrightarrow \frac{\|\Delta x\|}{\|A\| \cdot \|x\|} \leq \frac{\|A^{-1}\| \cdot \|\Delta b\|}{\|b\|}$$

hence

$$\frac{\|\Delta x\|}{\|x\|} \leq \underbrace{\|A\| \cdot \|A^{-1}\|}_{\text{cond}(A)} \frac{\|\Delta b\|}{\|b\|} \quad (3)$$

In fact, (3) is sharp:

let  $x$  s.t.  $\|Ax\| = \|A\| \cdot \|x\|$

$\Delta b$  s.t.  $\|A^{-1}\Delta b\| = \|A^{-1}\| \cdot \|\Delta b\|$

hence

$$\frac{(a)}{(b)} \Leftrightarrow \frac{\|\Delta x\|}{\|A\| \cdot \|x\|} \leq \frac{\|A^{-1}\| \cdot \|\Delta b\|}{\|b\|}$$

hence

$$\frac{\|\Delta x\|}{\|x\|} \leq \underbrace{\|A\| \cdot \|A^{-1}\|}_{\text{cond}(A)} \frac{\|\Delta b\|}{\|b\|} \quad (3)$$

In fact, (3) is sharp:

let  $x$  s.t.  $\|Ax\| = \|A\| \cdot \|x\|$

$\Delta b$  s.t.  $\|A^{-1}\Delta b\| = \|A^{-1}\| \cdot \|\Delta b\|$

## Theorem

$$\text{cond}(A) = \frac{\lambda_{\max}}{\lambda_{\min}}$$

when  $\lambda_{\min} \leq \dots \leq \lambda_{\max}$  are the  $n$  (real) eigenvalues of a symmetric matrix  $A$ .

Here:

$$\frac{\lambda_{\max}}{\lambda_{\min}} \approx 3000.$$

In (3) for

$$\|x\| = 2, \quad \|\Delta b\| = 0.1, \quad \|b\| \approx 60$$

$$\Rightarrow \|\Delta x\| \approx \frac{2 \times 3000 \cdot 0.1}{60} = 10$$

(cf. (1)) magnitude of error.

## Claim

$\lambda_1, \dots, \lambda_n$  are the eigenvalues of  $A^T A$ .

$$\|A\|^2 = \lambda_{\max} = \max\{\lambda_1, \dots, \lambda_n\}$$

Proof:

Let  $u_1, \dots, u_n$  the ON-basis of eigenvectors for  $A^T A$  with eigenvalues  $0 \leq \lambda_1 \leq \dots \leq \lambda_n$  ( $A^T A > 0$ ).

For any unit vector  $x = \sum_{i=1}^n c_i u_i$  we have  $\sum c_i^2 = 1$  and

$$\begin{aligned} \|A\|^2 &= \sup_{\|x\|=1} \langle Ax, Ax \rangle = \sup_{\|x\|=1} x^T \underbrace{A^T A}_{\sum c_i A^T A u_i = \sum c_i \lambda_i u_i} x \\ &= \sup_{\|\sum c_i u_i\|=1} \sum \lambda_i c_i^2 \leq \lambda_{\max} = \lambda_n \end{aligned}$$

Equality:  $c_n = 1, c_1 = 0, \dots, c_{n-1} = 0$ . □

This is a measure for *collinearity* in regression (redundant information).

$$\text{If } Y = X\beta + \epsilon, \quad \beta \in \mathbb{R}^m, \quad Y \in \mathbb{R}^n$$
$$X = [X_1, \dots, X_m]$$

Collinearity means:

$$\sum \lambda_j X_j \approx 0$$

then  $A = X^T X$  is close to be non-invertible.

- 1) Drop variables
- 2) Mean center the variables (or use other transformation, PCA!)
- 3) Orthogonalize  $X$  (this stabilises the variance)

$$1, x, x^2, \dots \quad \text{on } [0, 1]$$

→ shifted Legendre polynomials as "explanatory variables"

$$1, \sqrt{12} \left( x - \frac{1}{2} \right), \sqrt{5} \left( 6 \left( x - \frac{1}{2} \right)^2 - \frac{1}{2} \right), \dots \quad [\text{Abramowitz, Stegun, '72}]$$

- good from a numerical and statistical perspective
- However, interpretation of "new" Fourier coefficients may be difficult.

4) Regularise the problem:  $cond(A) = \frac{\lambda_{\max}}{\lambda_{\min}}$

e.g. by using an estimator, s.t.  $\lambda_{\min}$  does not become too small.  
(Ridge-regression)

What we have learned: Numerical stability (small condition number) roughly corresponds to statistical stability (small MSE)

4) Regularise the problem:  $cond(A) = \frac{\lambda_{\max}}{\lambda_{\min}}$

e.g. by using an estimator, s.t.  $\lambda_{\min}$  does not become too small.  
(Ridge-regression)

**What we have learned: Numerical stability (small condition number) roughly corresponds to statistical stability (small MSE)**