

Last time : 2.1 (b)

Matrix Calculus 101 (Jacobian convention)

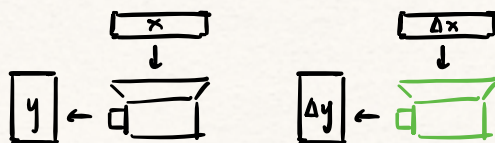
$$y = Ax \rightarrow \frac{\partial y}{\partial x} = A$$

$$y = x^T A \rightarrow \frac{\partial y}{\partial x} = A^T$$

$$y = x^T x \rightarrow \frac{\partial y}{\partial x} = 2x^T$$

$$y = x^T A x \rightarrow \frac{\partial y}{\partial x} = x^T (A + A^T) = 2x^T A \text{ if } A \text{ symmetric}$$

$$\begin{bmatrix} y \end{bmatrix} = \begin{bmatrix} A \end{bmatrix} \begin{bmatrix} x \end{bmatrix}$$



$$\text{Chain rule: } y = f(g(x)) \rightarrow \frac{\partial y}{\partial x} = \frac{\partial y}{\partial g} \frac{\partial g}{\partial x}$$

Approach 2: Matrix Calculus

$$\nabla_w L(w) = \nabla_w \|xw - y\|_2^2$$

$$= \nabla_w (xw - y)^T (xw - y)$$

$$= 2 (xw - y)^T x$$

by chain rule, now vector

$$\text{Set } \nabla_w L(w) \text{ to } 0: 2(xw - y)^T x = 0$$

$$2x^T (xw - y) = 0$$

reshape into column

$$2x^T xw - 2x^T y = 0$$

$$w = (x^T x)^{-1} x^T y$$

Linear Algebra Review (Cont'd)

Norms

$$\text{Let } v \in \mathbb{R}^n$$

$$1\text{-norm } (\ell_1): \|v\|_1 = \sum_{i=1}^n |v_i|$$

$$2\text{-norm } (\ell_2): \|v\|_2 = \sqrt{v^T v} = \sqrt{\sum_{i=1}^n v_i^2}$$

$$\infty\text{-norm } (\ell_\infty): \|v\|_\infty = \max_i |v_i|$$

↑ bigger = more sensitive to outliers

Matrices

Let $A \in \mathbb{R}^{n \times n}$

Symmetric : $A = A^T$

quadratic form : $x^T A x$ for $x \in \mathbb{R}^n$

positive semidefinite : $x^T A x \geq 0 \quad \forall x \in \mathbb{R}^n$

2.2

$$X = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

Row space : span (row vectors)

Column space (Range(X)) : span (column vectors)

Nullspace : $\{v : Xv = 0\}$

Rank : maximum number of L.I. rows / columns (they are equal)

2.3 (a) $\text{Row}(X) = \text{Col}(X^T)$

rows of X are columns of X^T

(b) nullspace of X and row space of X are orthogonal complements.

$$\text{Null}(X) = \text{Row}(X)^\perp$$

$$v \in \text{Null}(X) \Leftrightarrow Xv = \vec{0} \Leftrightarrow \langle X_i, v \rangle = 0 \quad \forall i$$

v perpendicular to all row vectors.

(c) $\text{Null}(X^T) = \text{Col}(X)^\perp$

$$\text{Null}(X^T) = \text{Row}(X^T)^\perp = \text{Col}(X)^\perp$$

Rank - Nullity Theorem : $A \in \mathbb{R}^{m \times n}$, $\text{Rank}(A) + \text{Nullity}(A) = n$

2.4 (a) $X^T X$ is symmetric and PSD

$$\text{Symmetric: } (X^T X)^T = X^T (X^T)^T = X^T X$$
$$\text{PSD: } v^T X^T X v = \|Xv\|_2^2 \geq 0$$

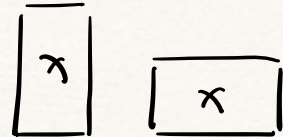
(b) $\text{Null}(X) = \text{Null}(X^T X)$

$$\text{Bidirectional proof: } v \in \text{Null}(X) \Rightarrow Xv = 0 \Rightarrow X^T Xv = 0$$
$$(X^T X)v = 0 \Rightarrow v^T X^T Xv = 0 \Rightarrow \|Xv\|_2^2 = 0$$
$$\Rightarrow Xv = 0$$

(c) $X^T X$ is not always invertible

use part (b)

(d) "tall skinny" vs "short wide" X



$$\text{We'll see: arg min}_w \|Y - Xw\|_2^2 = (X^T X)^{-1} X^T y$$

"short wide" $\rightarrow$ non empty null space $\rightarrow X^T X$ not invertible

$$\text{(e) } \text{Row}(X^T X) = \text{Col}((X^T X)^T) = \text{Col}(X^T X)$$

$$\text{Row}(X^T X) = \text{Null}(X^T X)^\perp = \text{Null}(X)^\perp = \text{Row}(X)$$

Maximum Likelihood Estimation

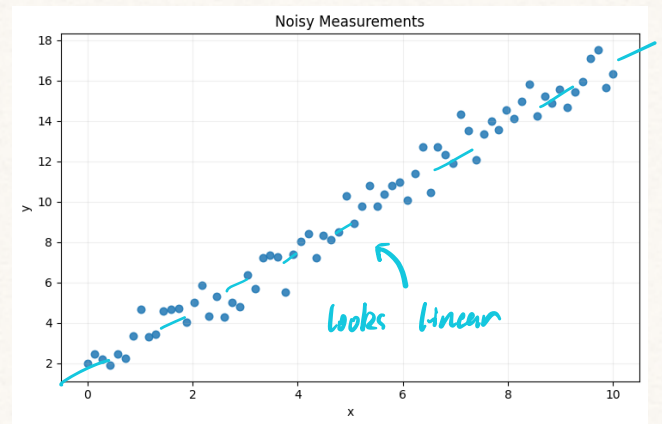
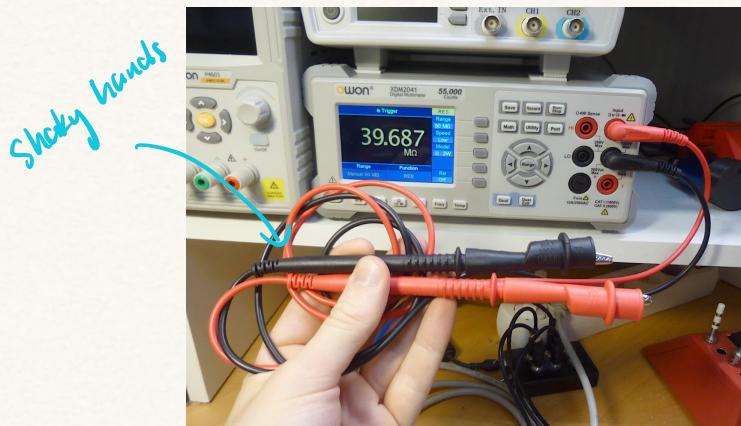
Given: 1. Data

2. Model (parametrized by θ)

Likelihood: $p_{\text{model}}(\text{Data} | \theta)$

Linear Regression

Say we have some resistors at different lengths
And we want to measure their resistance.



(a) Model: $y_i = x_i^T w + v_i$, $i \in 1, 2, \dots, m$
 $v_i \sim \mathcal{N}(0, \sigma^2)$

Goal: find $\arg\max_w \log p_w(x_1, y_1) \dots (x_m, y_m)$

Show: MLE is the solution to $\min_w \|Xw - Y\|_2^2$

$$\begin{aligned}\hat{w}_{MLE} &= \arg\max_w \log p_w((x_1, y_1) \dots (x_m, y_m)) \\ &= \arg\max_w \log \prod_{i=1}^m \mathcal{N}(y_i - x_i^T w; 0, \sigma^2) \\ &= \arg\max_w \sum_{i=1}^m \log \mathcal{N}(y_i - x_i^T w; 0, \sigma^2) \\ &= \arg\max_w \sum_{i=1}^m \log \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - x_i^T w)^2}{2\sigma^2}} \right) \\ &= \arg\max_w \cancel{m \log \frac{1}{\sqrt{2\pi}\sigma}} - \sum_{i=1}^m \frac{(y_i - x_i^T w)^2}{2\sigma^2} \\ &= \arg\max_w - \cancel{\frac{1}{2\sigma^2}} \sum_{i=1}^m (y_i - x_i^T w)^2 \\ &= \arg\min_w \sum_{i=1}^m (y_i - x_i^T w)^2 \\ &= \arg\min_w \|Xw - Y\|_2^2\end{aligned}$$

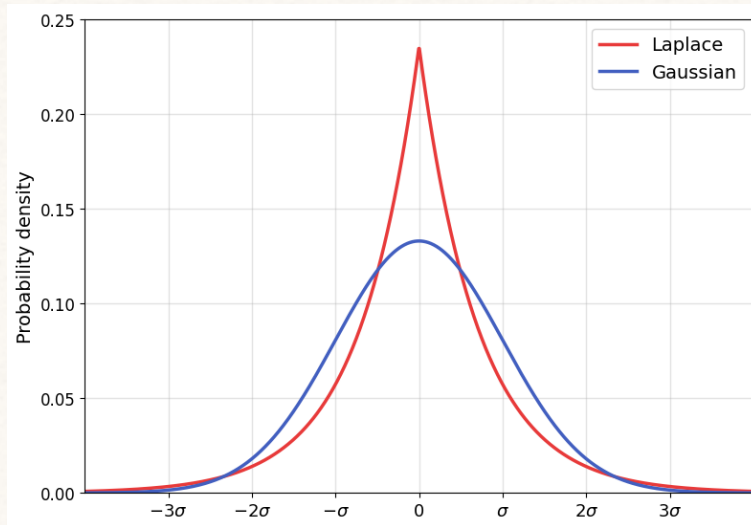
Next: plug in results from above:

$$\nabla \|Xw - Y\|_1^2 = 2X^T(Xw - Y)$$

set to 0:

$$\hat{w} = (X^T X)^{-1} X^T Y$$

(b) Can assume other noise distributions (e.g. hand tends to drop lower)



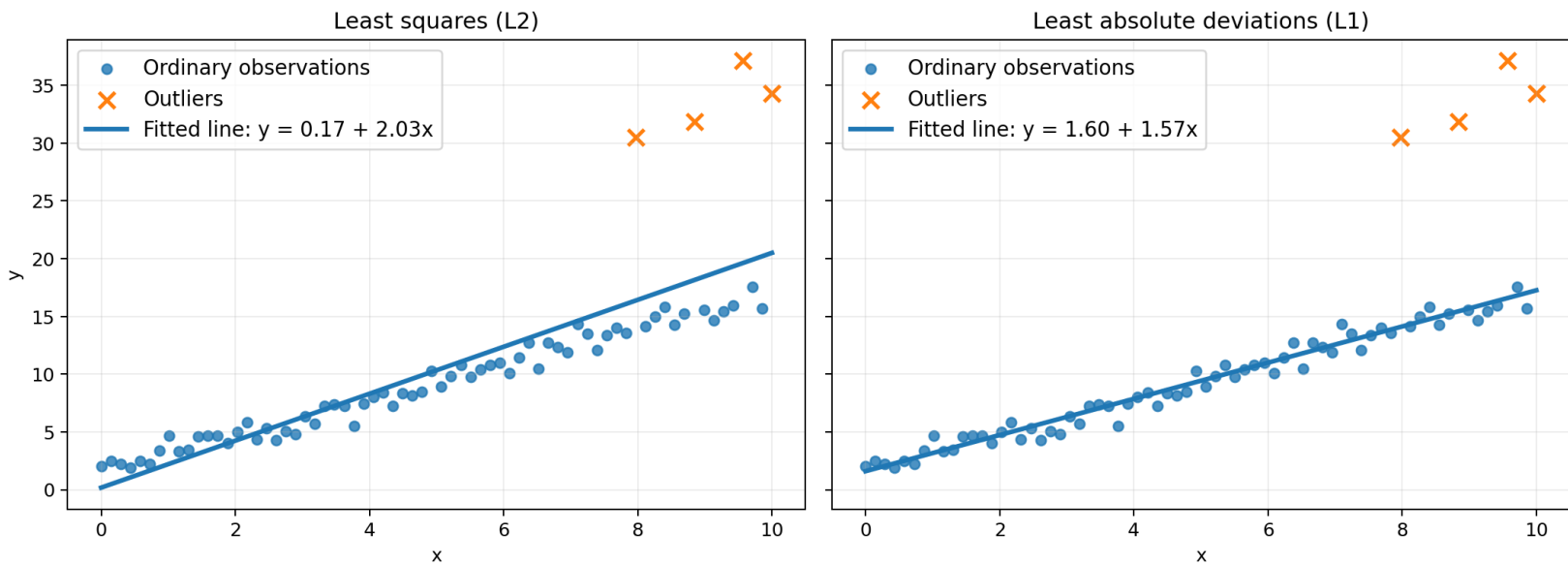
Model: $y_i = x_i^T w + v_i$, $i \in 1, 2, \dots, m$

$v_i \sim \text{Laplace}(0, b)$, $p(z) = \frac{1}{2b} e^{-\frac{|z|}{b}}$

Goal: find $\arg\max_w \log p_w(x_1, y_1) \dots (x_m, y_m)$

$$\begin{aligned} \hat{w}_{MLE} &= \arg\max \sum_{i=1}^m \log \text{Laplace}(y_i - x_i^T w; 0, b) \\ &= \arg\max \sum_{i=1}^m \log \left(\frac{1}{2b} e^{-\frac{|y_i - x_i^T w|}{b}} \right) \\ &= \arg\max m \log \frac{1}{2b} - \sum_{i=1}^m \frac{|y_i - x_i^T w|}{b} \\ &= \arg\max -\frac{1}{b} \sum_{i=1}^m |y_i - x_i^T w| \\ &= \arg\min \|Xw - Y\|_1 \end{aligned}$$

Intuition : Least Squares vs Least Absolute Deviation



LS : Gaussian - estimate the mean, sensitive to outliers

LAD : Laplacian - estimate the median, less sensitive

