Department of COMPUTER SCIENCE

WARWICK
THE UNIVERSITY OF WARWICK

Machine Learning
CS342

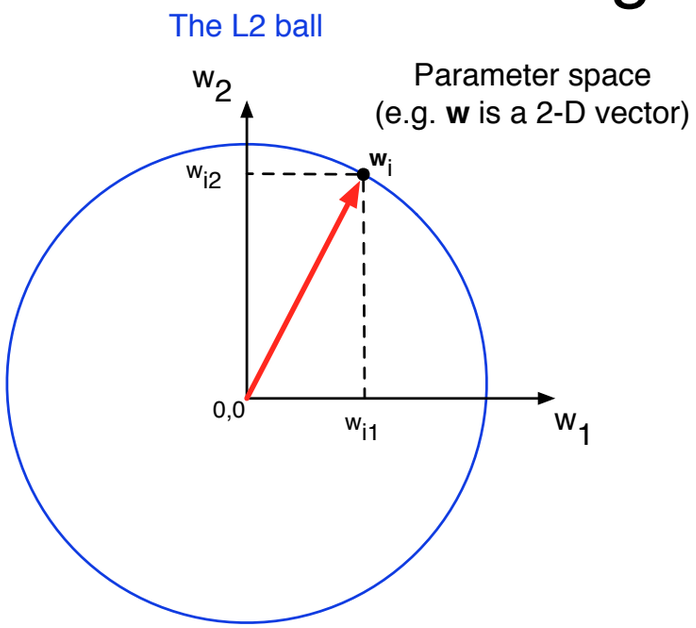
Lecture 8: The Maximum Likelihood framework

Dr. Theo Damoulas
T.Damoulas@warwick.ac.uk

Office hours (CS 307)
Mon 16:00-17:00 Fri 16:00-17:00

Recap: Regularised Linear regression (PLS vs Lasso)

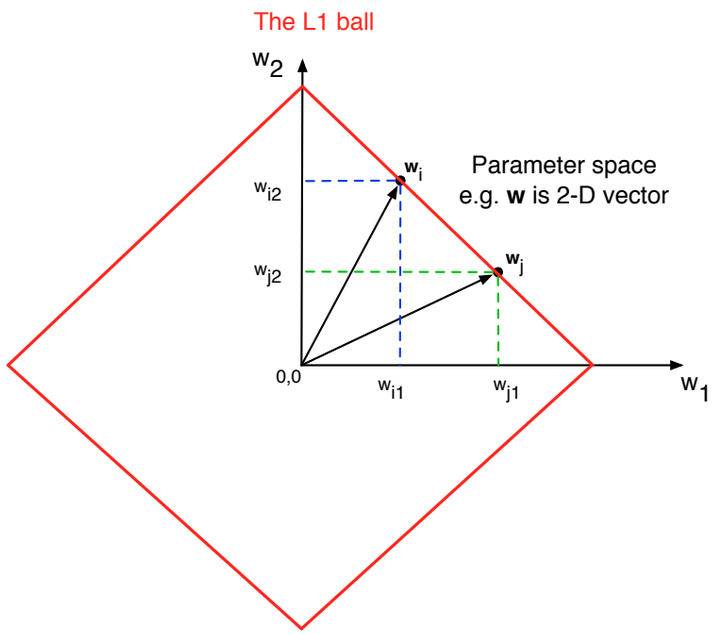
Regularisation to avoid overfitting in OLS



PLS/Ridge regression

$$\mathrm{L}_2^2(\mathbf{w}) = \sum_d w_d^2 \qquad \mathcal{L}' = \mathcal{L} + \lambda \mathbf{w}^{\mathrm{T}}\mathbf{w}$$

$$\text{Minimise } \mathcal{L} \qquad \text{s.t. } \sum_d w_d^2 = \mathbf{w}^{\mathrm{T}}\mathbf{w} \leq t$$



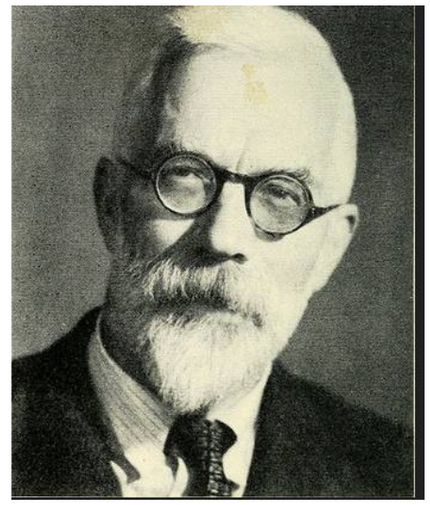
The Lasso

$$\mathrm{L}_1(\mathbf{w}) = \sum_{d=1}^{D} |w_d| \qquad \mathcal{L}' = \mathcal{L} + \lambda \sum_d |w_d|$$

$$\text{Minimise } \mathcal{L} \qquad \text{s.t. } \sum_d |w_d| \leq t$$

Maximum Likelihood: Errors as random noise

Statistical framework - not a model!
A way of thinking about ***"errors" as random variables***


Sir R. A. Fisher

"The Maximum Likelihood principle" - can be applied to all SL problems

We will study this principle in the context of a setting we understand:
Linear regression!

In this lecture we will end up with the **exact same solution as OLS**
but through The Maximum Likelihood principle

We will think ***Generatively:*** *How has our data been generated?*

Errors as Noise

Rogers & Girolami, Ch. 2

Requires familiarity with random variables and probability…
Support: R&G book 2.2.1 - 2.7 and module website material

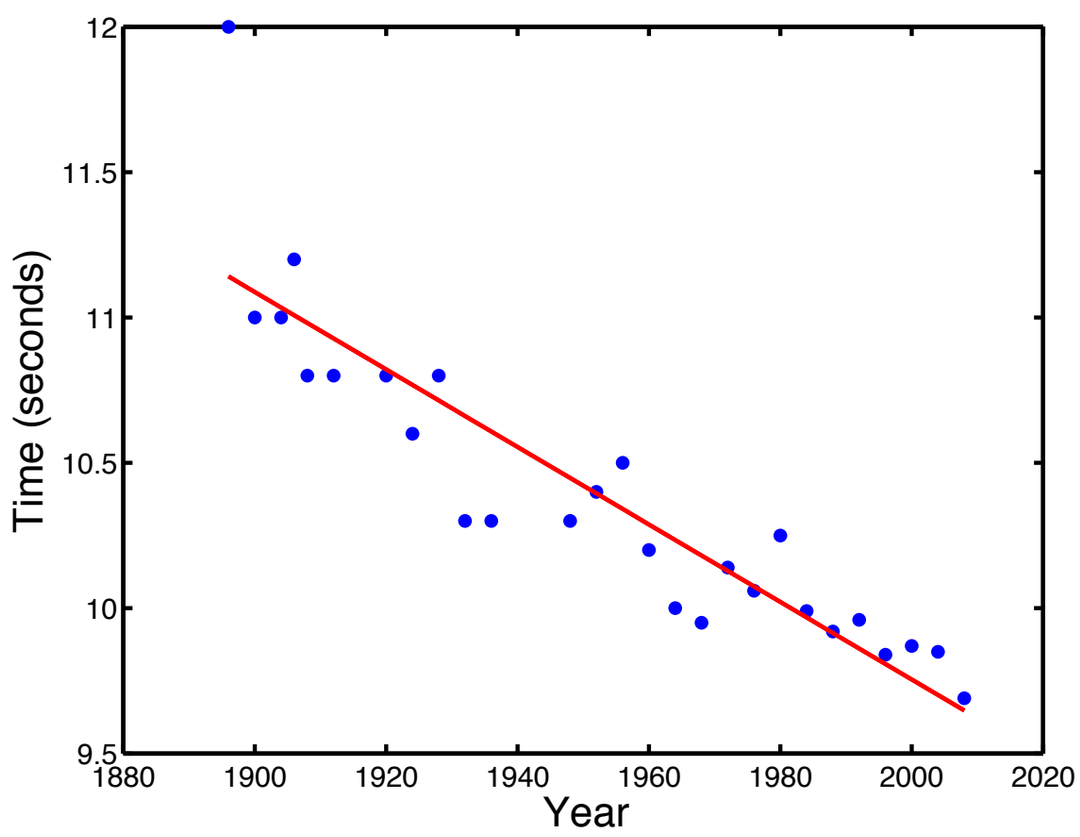
What was the “framework” we followed so far in LinReg (OLS/PLS/Lasso)?

Choose a Loss function (squared error), perhaps add regulariser.
Then Minimise them with respect to w to get final parameters/solution

Why linear hypothesis?

What are we saying about the underlying process that generated our data?

What is “noise”?



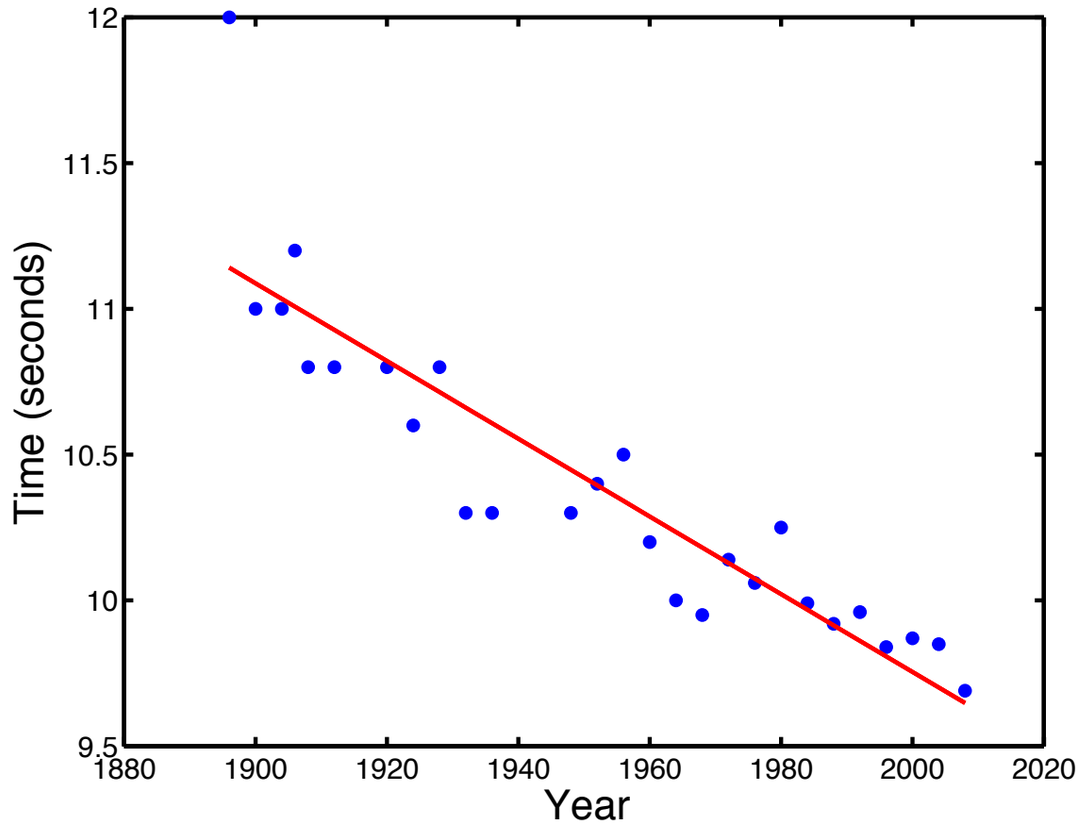
Errors as Noise

We will think ***Generatively:*** *How has our data been generated?*

Ok lets call it noise.
Is there an obvious pattern?

Looks deviations from line are random
Noise is “random”

What is “random”??



Theo’s working definition of random:
Anything that we lack the information and/or the computational capabilities in order to compute/predict.

Randomness

How do we compute/use random numbers in our computers?

Random number generators

Wikipedia: A **random number generator** (**RNG**) is a computational or physical device designed to generate a sequence of numbers or symbols that can not be reasonably predicted better than by a random chance

Pretty circular… lets summarise what an RNG does

RNG: Produces “pseudo-random” numbers based on increasingly complex ***patterns***

Food for thought…

Department of COMPUTER SCIENCE

Randomness

Entropy: A measure of structure/order/homogeneity

Out of the box: High Entropy
(very “random”)

As we build it we reduce Entropy.
(less “random”)

Ok enough with “philosophy” lets go back to linear regression

Random variables 101

- A discrete random variable has a Probability Distribution Function (PDF)
- e.g. Rolling a dice (discrete events)

$$0 \leq P(X = x) \leq 1 \qquad \sum_x P(X = x) = 1$$

- What is the expected value of rolling a fair dice?

$$\mathbb{E}_{P(X)} = \sum_x xP(x) = ?$$

- A continuous random variable has a probability density function (pdf)

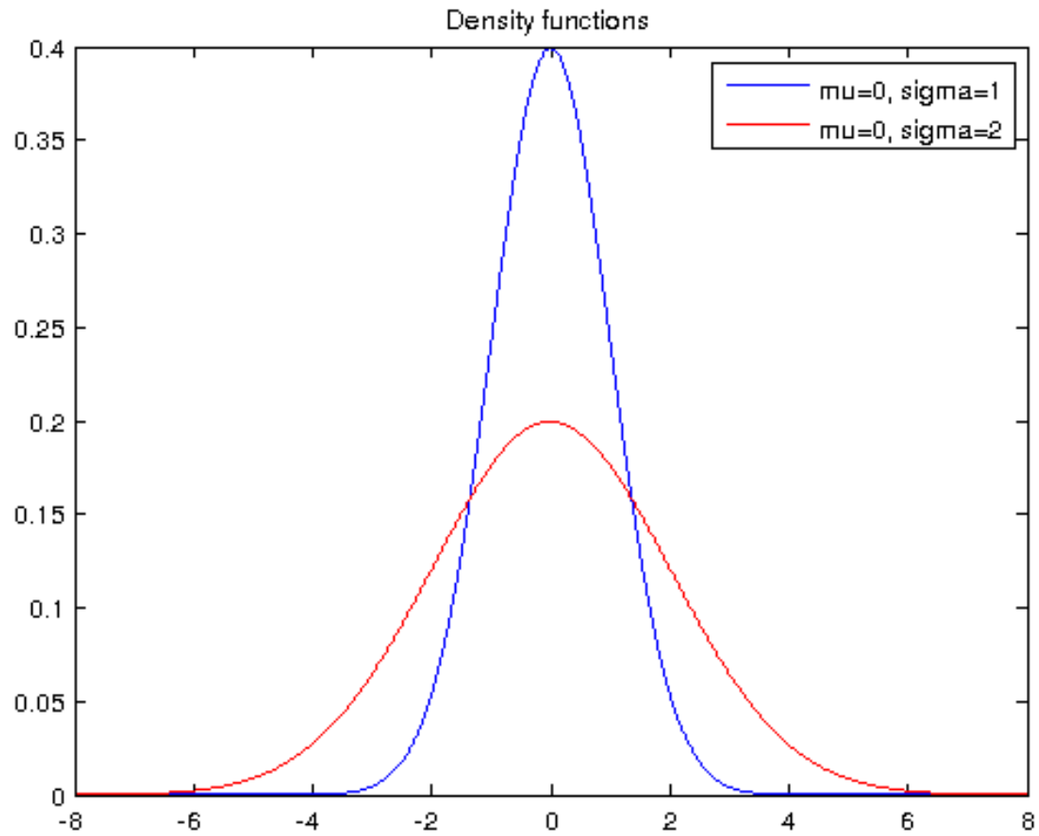
e.g. The Normal or Gaussian distribution

$$p(x) \sim \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Random variables 101

Gaussian white noise: 0-mean Normal/Gaussian distribution

$$p(x) \sim \mathcal{N}(0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2}\right\}$$



What is the expected value of x? Is the variance an expectation?

Department of COMPUTER SCIENCE

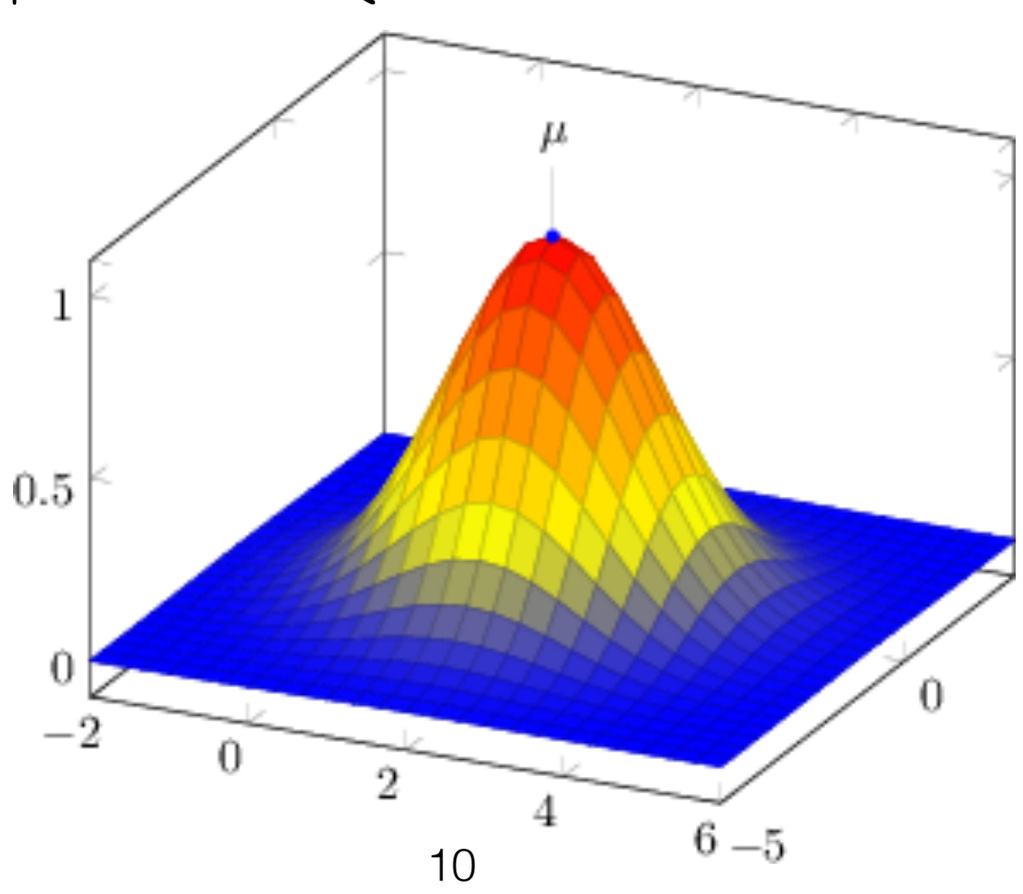
Random variables 101

So far X was a scalar. Can we place distributions over vectors?

Higher-dimensional space so distributions become also higher-D

So a “Multivariate Gaussian distribution” is the generalisation to higher-D

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\}$$



University of Warwick Computer Science Department

Initial Module Evaluation Form

The aim of this questionnaire is to **quickly** ascertain whether the module is running properly in its early stages. The results of the form will be viewed by the module organiser and their peer observation colleague so that they can make any necessary changes.

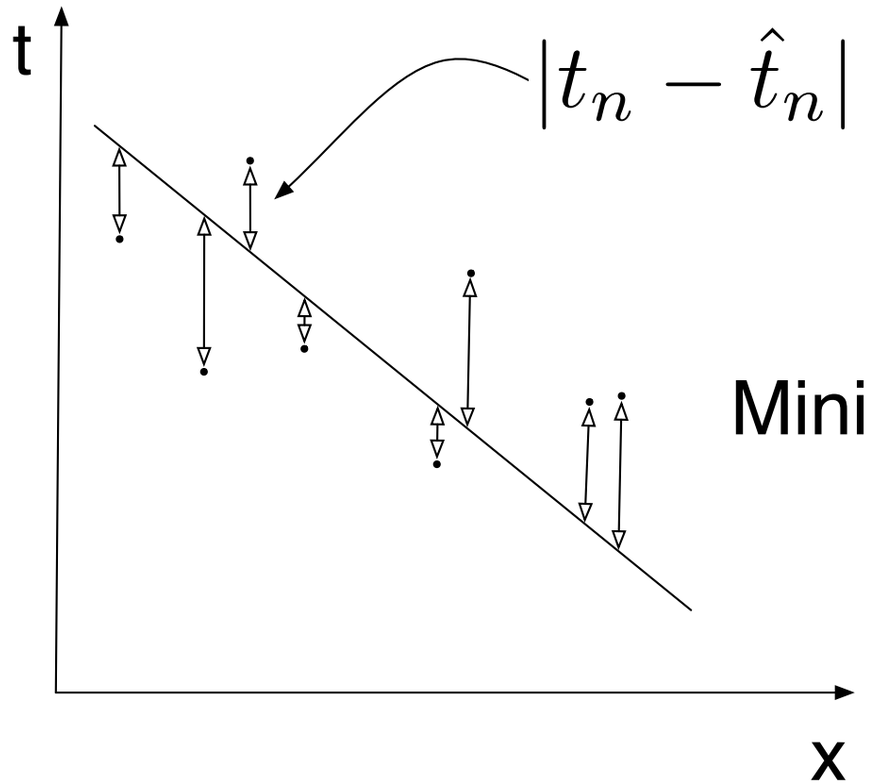
Please put a small mark (X) in the appropriate box. The results will be tallied by the organiser.

Module Code: Lecturer:

Can you hear the lecturer?		
Yes: 80/81	No: 1/81	
Can you read the lecturer's handwriting/Presentation Slides?		
Yes: 80/81	No: 1/81	
Is the rate of delivery...		
Too fast: 16/81	About right: 59/81	Too Slow: 6/81
Do the lectures seem well organised?		
Yes: 80/81	No: 1/81	
Has the lecturer made the academic objectives of the module clear?		
Yes: 78/81	No: 3/81	
Do you feel that your understanding is a sufficient grounding for this module?		
Yes: 61/81	No: 20/81	
Have you been informed of any relevant textbooks?		
Yes: 81/81	No: 0/81	
Have you been made aware of the assessment methods for this module?		
Yes: 81/81	No: 0/81	
Are you attending the support classes for this module (if provided)?		
Yes: 78/81	No: 3/81	

Errors as Noise

$$|t_n - \hat{t}_n| = |t_n - f(x_n; \mathbf{w})|$$

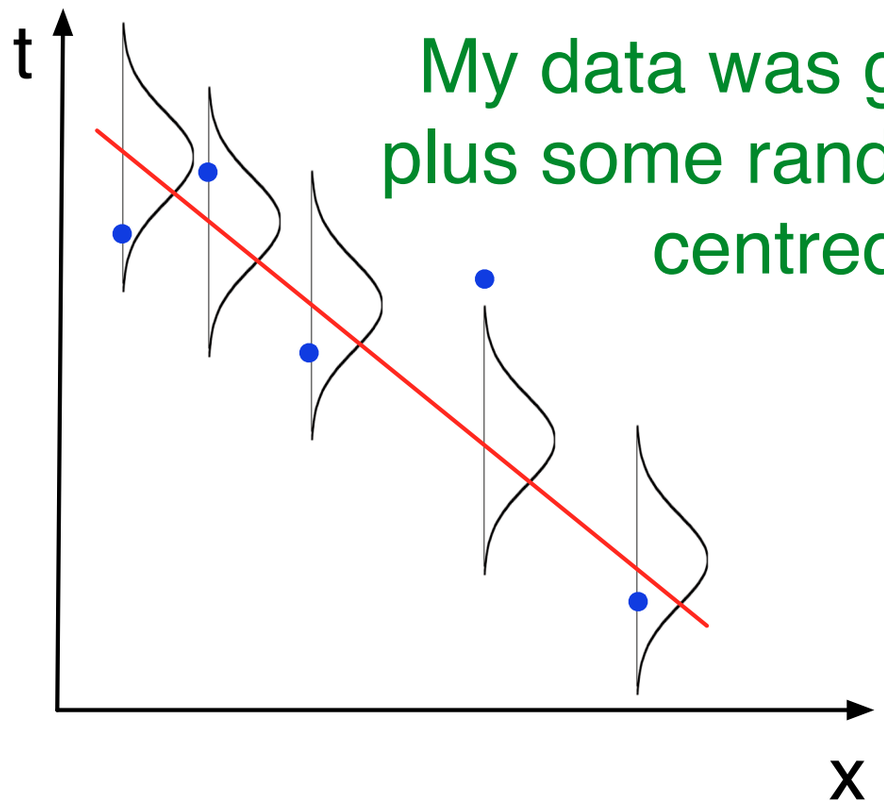


Error

OLS:
Minimise squared error

Think Generatively!

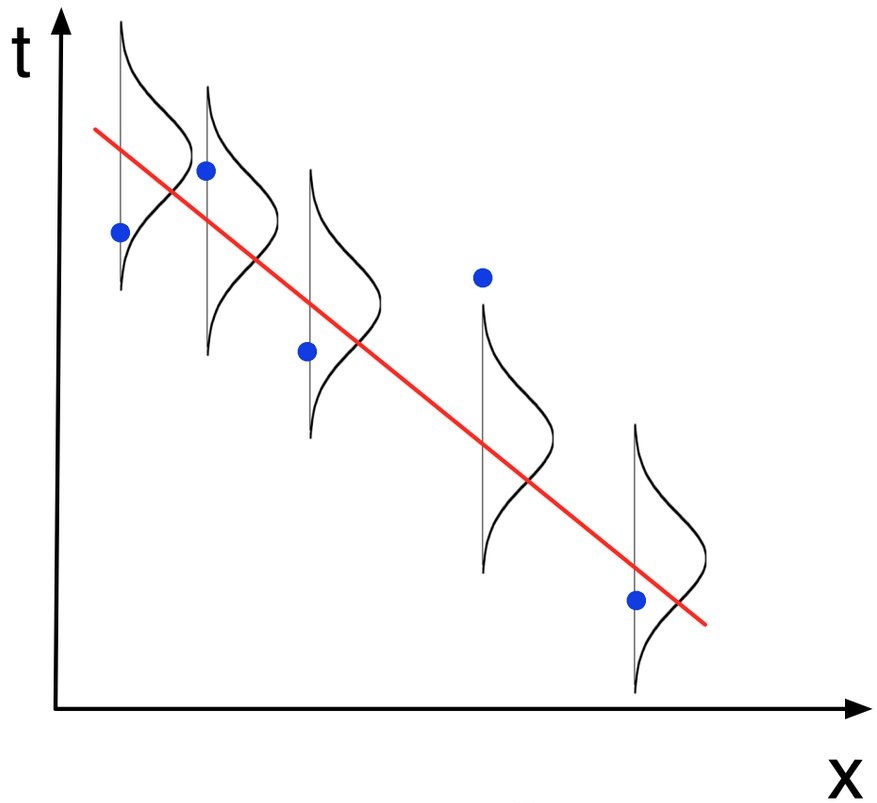
White Gaussian Noise



My data was generated by a line plus some random Gaussian noise centred on the line

Department of COMPUTER SCIENCE

Noise and Likelihood



So my model of what happened is:

$$t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n, \qquad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$

"My data was generated from a line (or plane in higher-D) plus some noise"

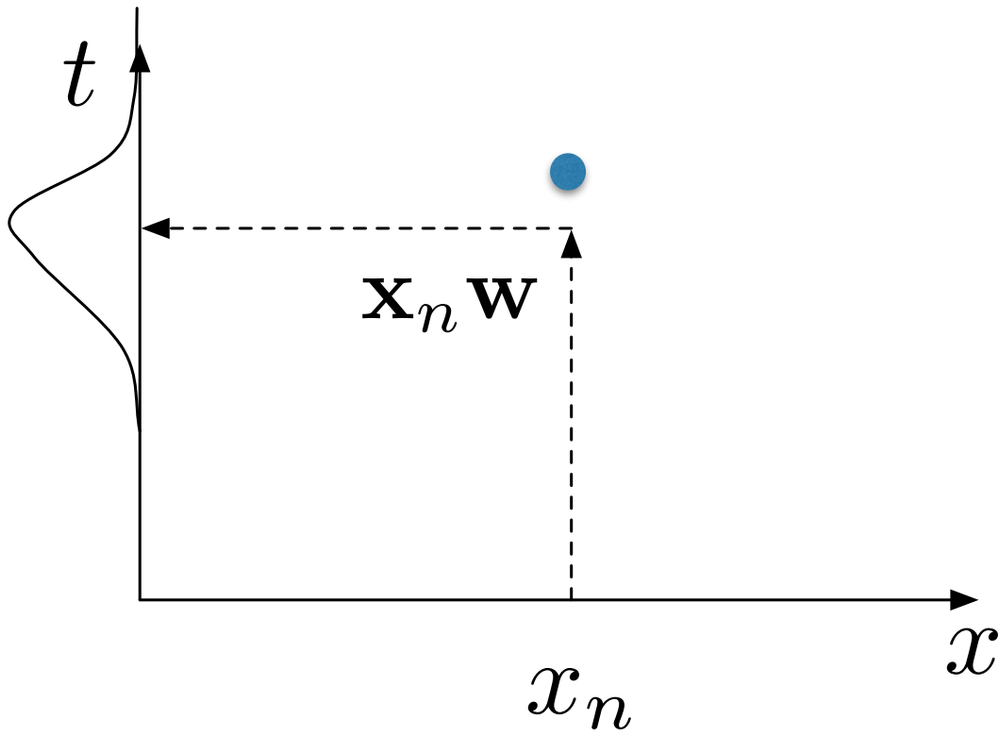
$$t_n = \mathbf{x}_n \mathbf{w} + \epsilon_n$$

deterministic component
(a.k.a. trend or drift)

random component
(a.k.a. noise term)

Noise and Likelihood

- Create your own synthetic data:
  - Create a line (signal):
    - Fix some **w**
    - sample some x, e.g. x~uniform
  - Add some noise:
    - For every point, $\mathbf{x}_n\mathbf{w}$, add Gaussian noise



Noise and Likelihood

Assumption: noise values are independent and homoscedastic

$$p(\epsilon_1, \ldots, \epsilon_N) = \prod_{n=1}^{N} p(\epsilon_n) = \prod_{n=1}^{N} \mathcal{N}(0, \sigma^2)$$

why independence gives us a product?

Substitute for noise term $t_n = \mathbf{x}_n \mathbf{w} + \mathcal{N}(0, \sigma^2)$

What happens when adding a constant to a normal distribution?

$$t_n \sim \mathcal{N}(\mathbf{x}_n \mathbf{w}, \sigma^2) \quad \text{so} \quad p(t_n | \mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{x}_n \mathbf{w}, \sigma^2)$$

Department of COMPUTER SCIENCE

Noise and Likelihood

$$p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{x}_n\mathbf{w}, \sigma^2)$$

Using the independence assumption to talk about all the data:

Likelihood

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n\mathbf{w}, \sigma^2)$$

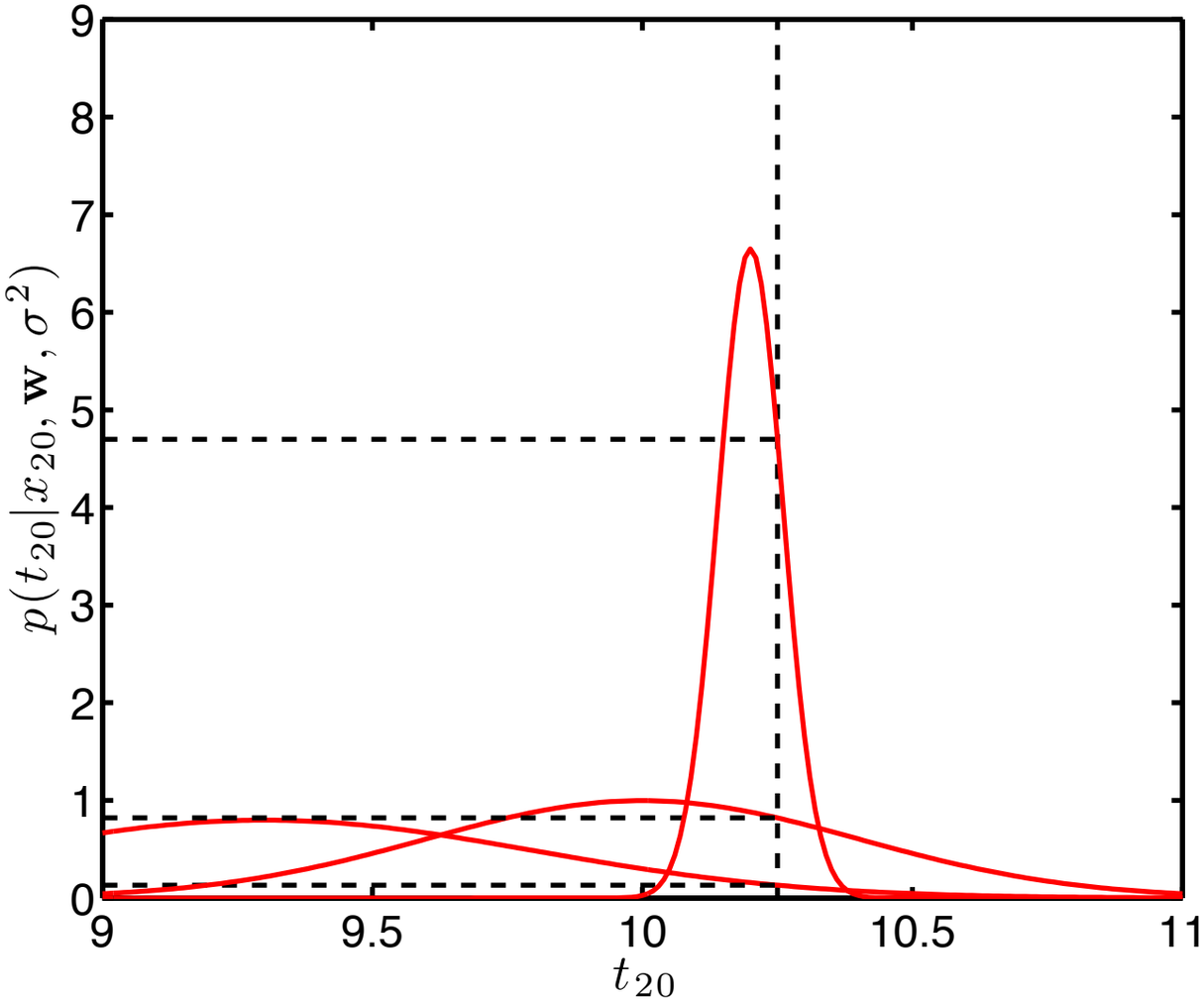
It is a function of the parameters

$$L(\mathbf{\Theta}) \text{ in this case : } L(\mathbf{w}, \sigma^2)$$

“How likely is that my model, with these parameters, can generate the data?

Likelihood: Olympic data

A single observation under different Gaussian likelihoods (different w,σ)

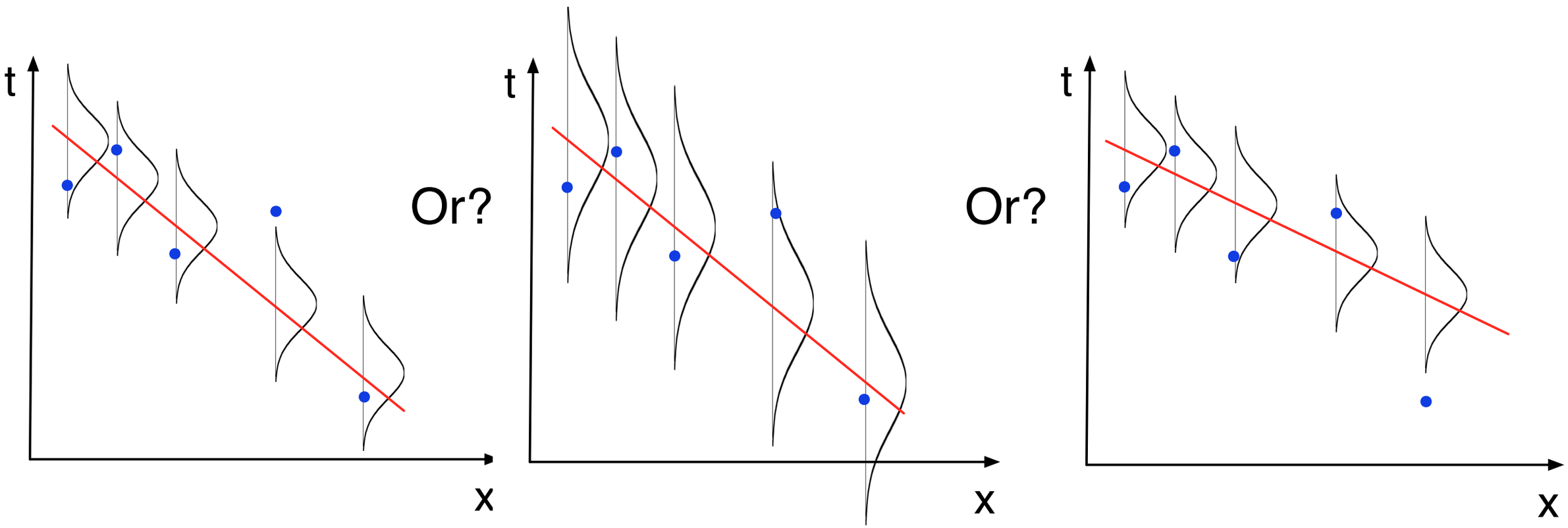


Which one looks better?

Likelihood: Many observations

Likelihood

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n \mathbf{w}, \sigma^2)$$



I can “tune” w, σ to get different likelihood - what do I want?

Department of COMPUTER SCIENCE

Maximum Likelihood

Find the parameters (w, σ) that maximise the likelihood!

$$\mathbf{w}, \sigma \leftarrow \text{argmax} \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n \mathbf{w}, \sigma^2)$$

- In fact we will maximise the (natural) log of this for convenience
- Similarly to OLS procedure, set 1st derivative to 0, check 2nd derivative

Maximum Likelihood

Substituting for the specific likelihood (Gaussian)

$$\mathbf{w}, \sigma \leftarrow \underset{\mathbf{w},\sigma}{\operatorname{argmax}} \sum_{n=1}^{N} \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(t_n - \mathbf{x}_n \mathbf{w})^2}{2\sigma^2} \right\} \right\}$$

We will end up with a term that looks like the sum of squared errors again

Convince yourself: $\boxed{\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{t}}$

Same solution as OLS! But we also learn the variance (noise model)

Maximum Likelihood

The maximum likelihood estimate for the variance is:

See book 2.8

$$\hat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mathbf{x}_n\hat{\mathbf{w}})^2$$

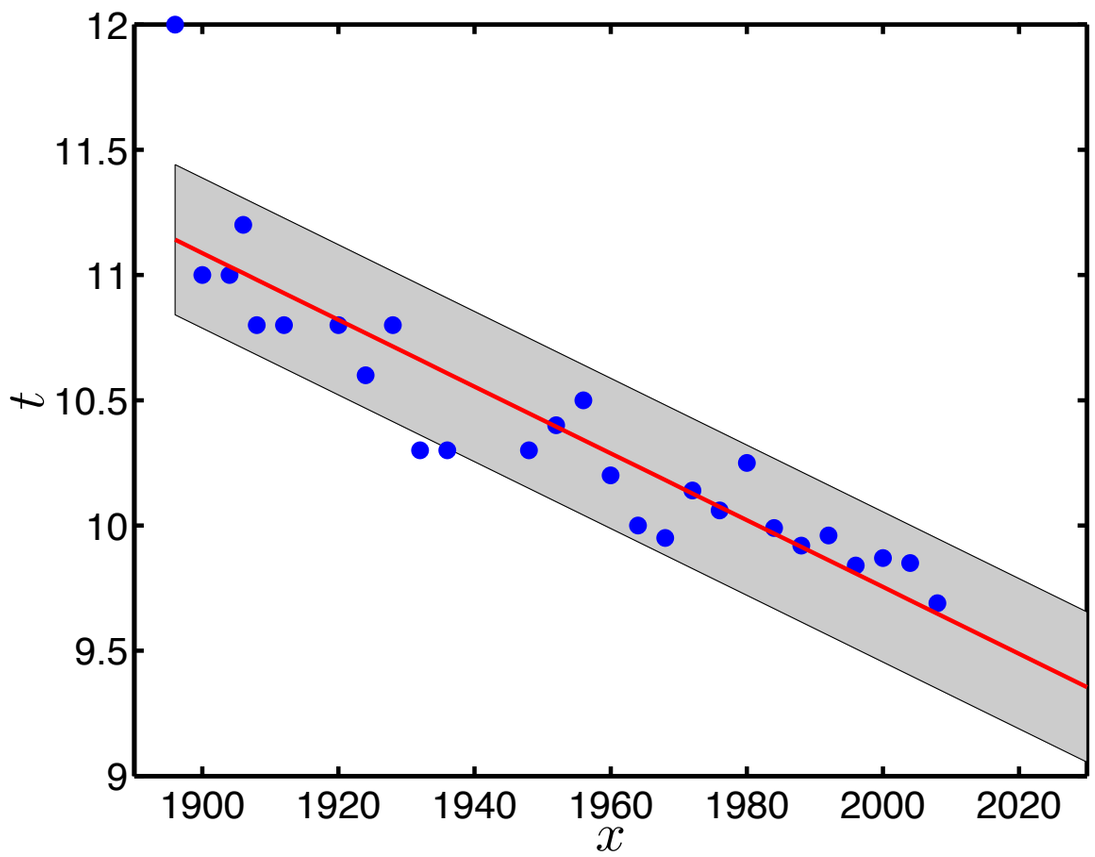
which is simply the average squared error!

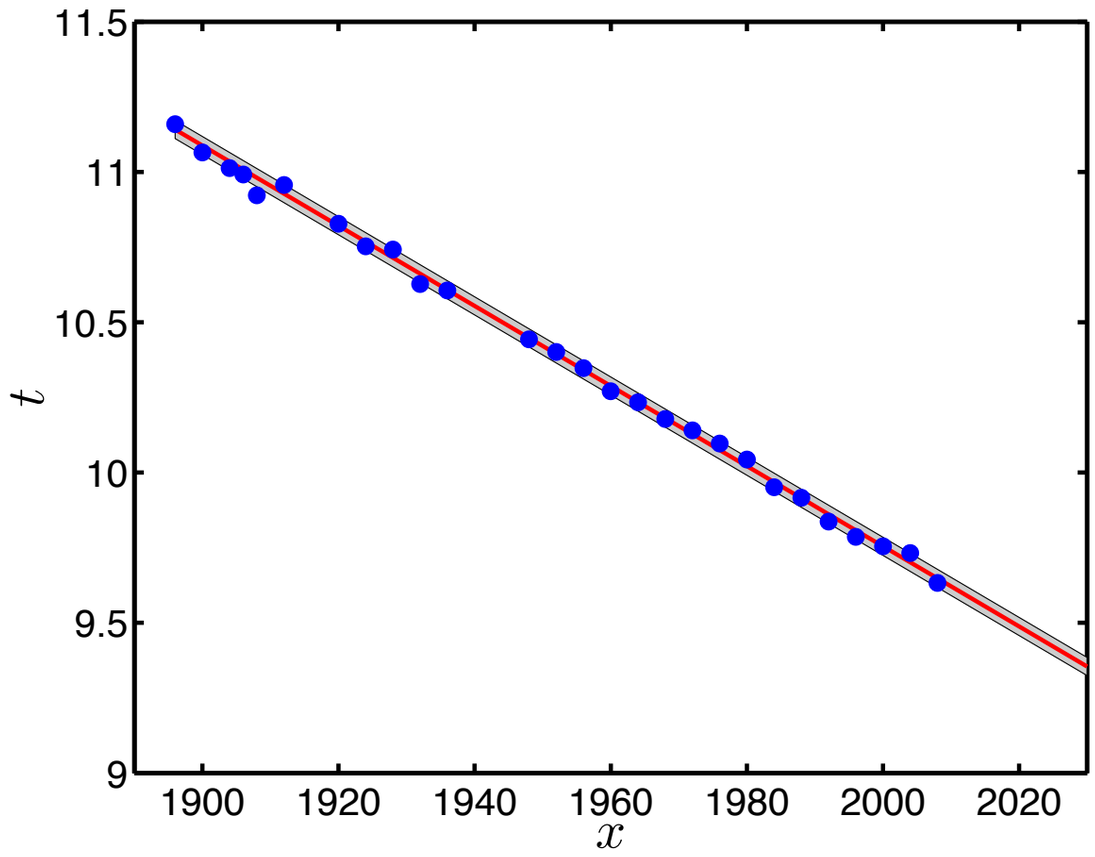
re-written in vector-matrix format:

$$\hat{\sigma^2} = \frac{1}{N}(\mathbf{t}^{\mathrm{T}}\mathbf{t} - \mathbf{t}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}})$$

Maximum Likelihood

What is the benefit of estimating the noise variance?





Summary: Maximum Likelihood

- Think Generatively
- Errors as Noise hence a random variable
- Linear model = deterministic + random noise component
- Via the definition of the noise model we end up with a Likelihood
- We maximise the likelihood to tune/find the parameters
- Maximum Likelihood solution for Lin. Reg. = OLS for w
- Similar problems with OLS (prone to overfitting, outliers)