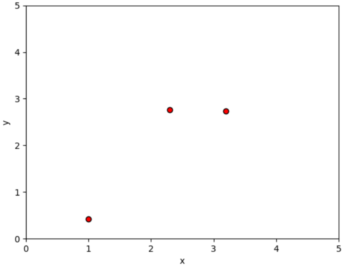
A gentle introduction to data assimilation

<u>Jérôme Fehrenbach</u>, Eliza Gyulgyulyan

Yerevan - june 2024
EASIM School

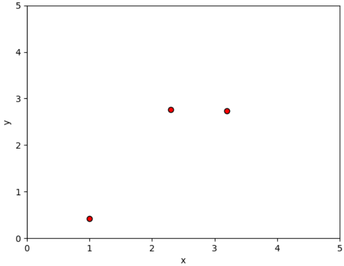
1) A very simple example
2) What is data assimilation ?
3) Another example
4) A global (subjective) view

3 data points in XY plane



Linear model:

$$d = Gm$$

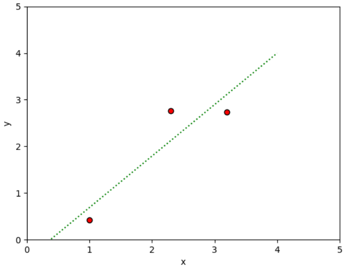


$Y = aX + b$
to be determined: $m = (a, b)^\top$

with $d = (y_1, y_2, y_3)^\top$ and $G = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \end{pmatrix}$

Linear model:

$$d = Gm$$



the overdertermined system

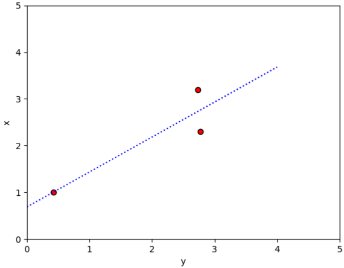
$$d = Gm$$

has no solution. We solve instead

$$m = (G^T G)^{-1} G^T d.$$

Linear model:

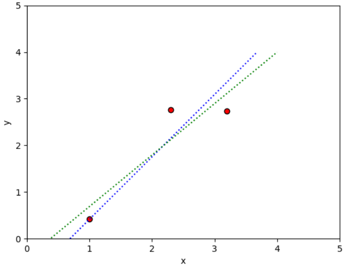
$$d = Gm$$



If x and y are exchanged

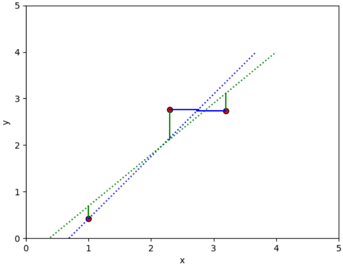
Linear model:

$$d = Gm$$



Linear model:

$$d = Gm$$



$d = Gm$

$$m = \underset{m}{\operatorname{argmin}} \sum_i (d_i - G_i(m))^2$$

$d = Gm$

$$m = \underset{m}{\operatorname{argmin}} \sum_i (d_i - G_i(m))^2$$

Now we propose a point of view that can be generalized.
for every i we assume that

$$d_i = G_i(m) + \epsilon_i,$$

where the noise satisfies $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

$d = Gm$

$$m = \underset{m}{\operatorname{argmin}} \sum_i (d_i - G_i(m))^2$$

Now we propose a point of view that can be generalized.
for every i we assume that

$$d_i = G_i(m) + \epsilon_i,$$

where the noise satisfies $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

Assume that the model m is known. What is the probability density to observe $(d_i)_i$?

$$p(d|m) = \prod_i p(d_i) \propto \exp\left(-\frac{\sum_i \epsilon_i^2}{2\sigma^2}\right) = \exp\left(-\frac{\sum_i (d_i - G_i(m))^2}{2\sigma^2}\right).$$

$d = Gm$

$$m = \underset{m}{\text{argmin}} \sum_i (d_i - G_i(m))^2$$

Now we propose a point of view that can be generalized.
for every i we assume that

$$d_i = G_i(m) + \epsilon_i,$$

where the noise satisfies $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

Assume that the model m is known. What is the probability density to observe $(d_i)_i$?

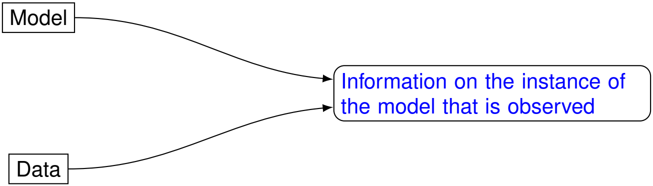
$$p(d|m) = \prod_i p(d_i) \propto \exp\left(-\frac{\sum_i \epsilon_i^2}{2\sigma^2}\right) = \exp\left(-\frac{\sum_i (d_i - G_i(m))^2}{2\sigma^2}\right).$$

The estimated model is the one that provides the most probable observations.

$$m = \text{argmax}\, p(d|m)$$

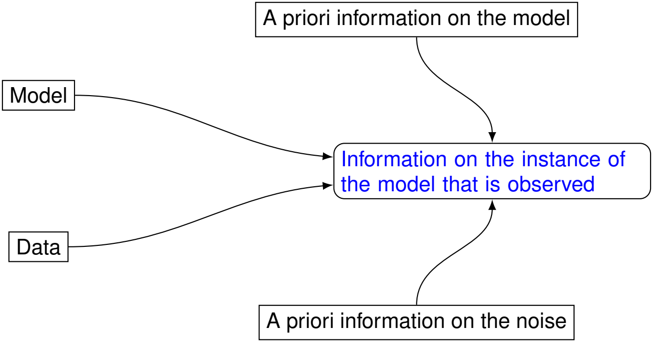
Maximize the likelihood $\equiv$ minimize $-$ the log-likelihood

Data assimilation consists in extracting information on the model from the observation of data.



Model: PDE, ODE, equation, ...
Data: measurement of some output of the model, possibly corrupted by noise.

Data assimilation consists in extracting information on the model from the observation of data.



The information provided by data assimilation can be

- are the parameters identifiable ?
- parameters estimation
- uncertainty on the parameters
- probability distribution of the parameters

The information provided by data assimilation can be

- are the parameters identifiable ?
- parameters estimation
- uncertainty on the parameters
- probability distribution of the parameters

The art of data assimilation consists in making use of a-priori information that is available.
This a-priori information can be

- a-priori probability distribution of the model parameters (Bayesian estimation)
- a-priori information on the regularity of the solution (inverse problem and regularization theory)
- information on the noise

I feel sick and I go to the doctor to pass a test for a disease.
The test is positive.

Question: am I affected by the disease ?

A-priori information:
"the test is not perfect, it provides 1% false positive and 1% false negative"

- Frequentist approach (data driven):

$$m = \text{argmax}\, p(d|m)$$

p = frequency distribution of a random process

$p(d|m = 0) = 0.01$, and $p(d|m = 1) = 0.99$.
The frequentist answer is : Yes I am sick.

• Frequentist approach (data driven):

$$m = \operatorname{argmax} p(d|m)$$

p = frequency distribution of a random process

$p(d|m=0) = 0.01$, and $p(d|m=1) = 0.99$.
The frequentist answer is : Yes I am sick.

• Bayesian approach: incorporate a-priori information on the model.
In our case: "1% of the population has the disease".

How to account for this information ?

$$m = \operatorname{argmax} p(m|d)$$

p=tool to describe the level of knowledge of an unknown, unobserved variable

Bayes' formula:

$$p(m|d) = \frac{p(d|m)p(m)}{p(d)}.$$

let us evaluate the probability of each alternative:

$$p(m=0|d) = \frac{p(d|m=0)p(m=0)}{p(d)} \propto 0.01 \times 0.99.$$

$$p(m=1|d) = \frac{p(d|m=1)p(m=1)}{p(d)} \propto 0.01 \times 0.99.$$

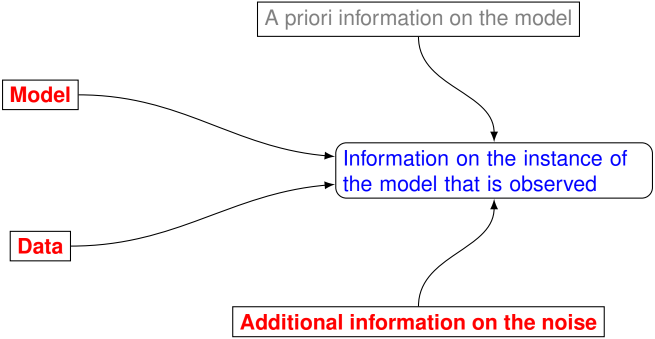
The Bayesian answer is: there is 50% chance that I am sick.

Information obtained	**Methods**	**Pros and cons**
proba. dist. of the parameters	MCMC, ...	✓ precise answer ✗ heavy computations
most probable param. (max. likelihood)	Variational data assim. Sequential data assim.	✓ relatively fast answer ✗ requires identifiability

Information obtained	**Methods**	**Pros and cons**
proba. dist. of the parameters	MCMC, ...	✓ precise answer ✗ heavy computations
most probable param. (max. likelihood)	**Variational data assim.** Sequential data assim.	✓ relatively fast answer ✗ requires identifiability

Variational means that we will **minimize** cost functions.

Course 1:


A priori information on the model
Model
Information on the instance of the model that is observed
Data
Additional information on the noise

- Day 1, part 1. Introduction
- Day 1, part 2. Optimization basics
- Day 2, part 1. Derivation of the tangent model
- Day 2, part 2. Derivation of the adjoint model