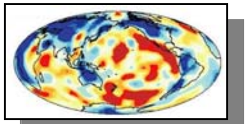
$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$



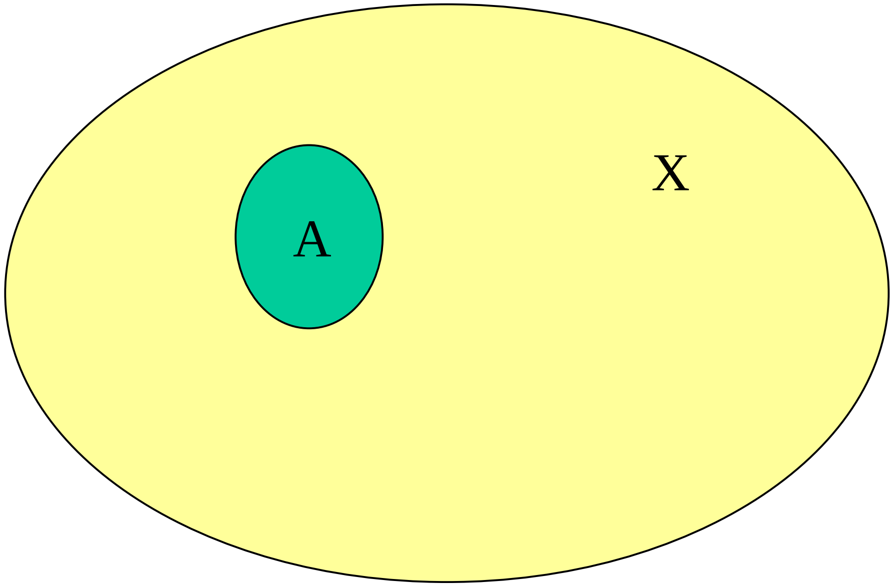
Probability and information

- The concept of probability
- probability density functions (pdf)
- Bayes' theorem
- states of information
- Shannon's information content
- Combining states of information
- The solution to inverse problems

Albert Tarantola

This lecture follows **Tarantola**, Inverse problem theory, p. 1-88.

Measures and Sets



Let X represent an arbitrary set.

What is a measure over X?

A measure over X implies that to any subset A of X a real positive Number P(A) is associated with the properties:

a. If Ø is the empty set then P(Ø) = 0.
b. If A_1, A_2, ... Are disjoint sequences of X then

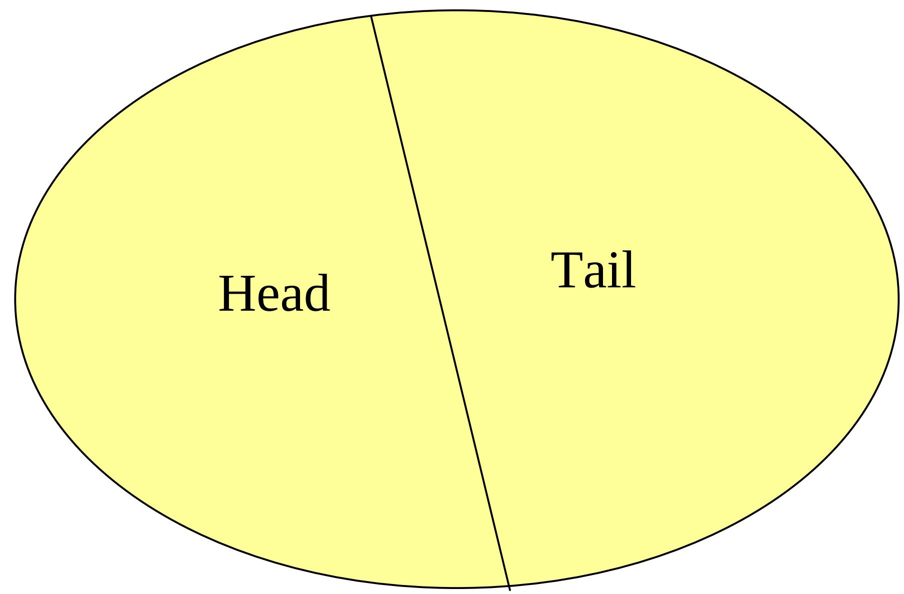
$$P\left[\sum_i A_i\right] = \sum_i P(A_i)$$

$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

Measures and Sets

P(X) is not necessarily finite.
If it is then we may call P a
probability or a probabilty measure.

P is usually normalized to unity.

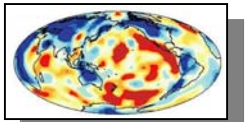


Example: Let X be {head,tail}

P(Ø)=P(neither head nor tail) = 0
P(head)=r
P(tail)=1-r

And P(head U tail) = 1

Probability density functions



As you expected we need to generalize this concept to continuous functions. In Earth sciences we often have functions of space coordinates such as f(x,y,z) and/or further variables $f(x_1,x_2,x_3, \ldots)$ If these functions exist such that for

$$A \subset X$$

$$P(A) = \int_A dx f(x)$$

$$\int_A dx = \int_A dx_1 \int_A dx_2 \int_A dx_3 \ldots$$

… then f(x) is termed a measure density function. If P is finite then f(x) us termed a **probability density function** … often called pdf .

Examples?
What are the physical dimensions of a pdf?

$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

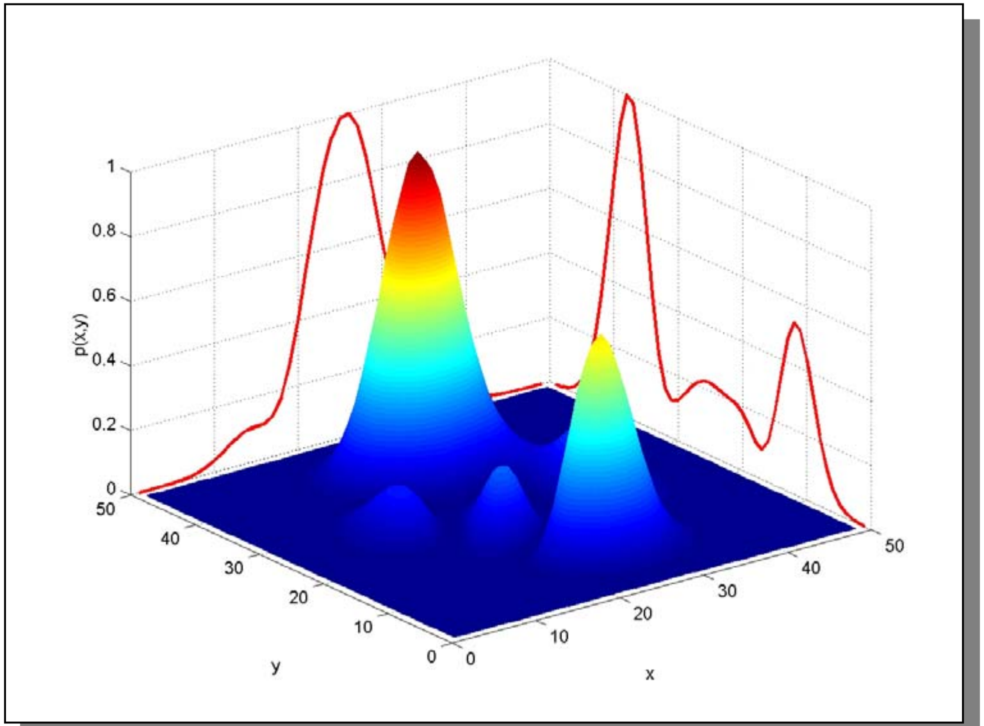
Marginal probabilities

Let x and y be two vector parameter sets

Example: x_i describes the seismic velocity model
y_i describes the density model

The **marginal probability density** is defined as

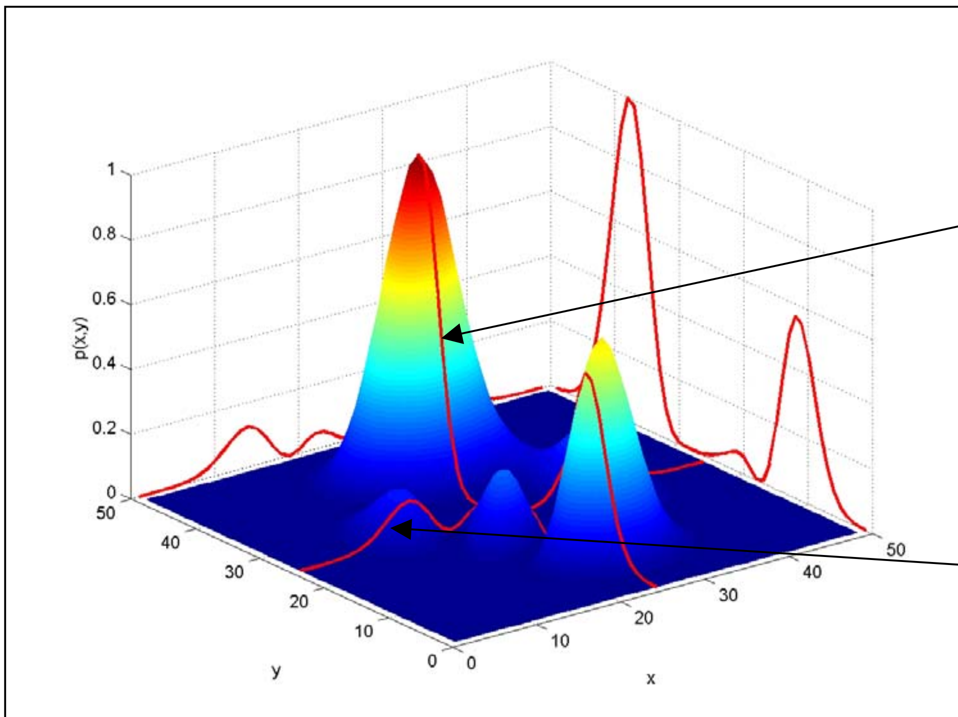
$$f_Y(y) = \int_X dx f(x,y)$$



$$\sigma(d,m)=k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

Conditional probabilities

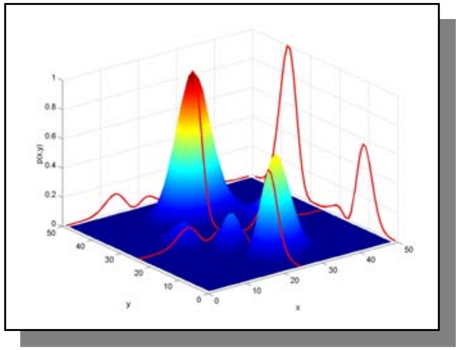
And the **conditional probability density** for x given $y=y_0$ is defined as



$$f_{Y|X}(y\,|\,x_0)=\frac{f(y,x_0)}{\int\limits_Y dxf(y,x_0)}$$

$$f_{X|Y}(x\,|\,y_0)=\frac{f(x,y_0)}{\int\limits_X dxf(x,y_0)}$$

Bayes Theorem



… it follows that, the joint pdf f(x,y) equals the conditional probability density times the marginal probability density

$$f(x,y) = f_{X|Y}(x \mid y)\, f_Y(y)$$

or

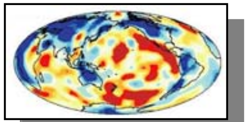
$$f(x,y) = f_{Y|X}(y \mid x)\, f_X(x)$$

Bayes theorem gives the probability for event y to happen given event x

$$\boxed{f_{Y|X}(y,x) = \frac{f_{X|Y}(x \mid y)\, f_Y(y)}{\int_Y f_{X|Y}(x \mid y)\, f_Y(y)dy}}$$

$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

The interpretation of probability



Possible interpretations of probability theory (Tarantola, 1988):

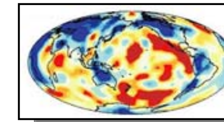
1. A purely statistical interpretation: probabilities diescribe the outcome of random processes (in physics, economics, biology, etc.)

2. Probabilities describe **subjective degree of knowledge** of the true value of a physical parameter. Subjective means that the knowledge gained on a physical system may vary from experiment to experiment.

The key postulate of **probabilistic inverse theory** is (Tarantola 1988):

Let X be a discrete parameter space with a finite number of parameters. The most general way we have for describing any state of information on X is by defining a probability (in general a measure) over X.

State of information



… to be more formal …

Let P denote the probability for a given state of information on the parameter space X and f(x) is the probability density

$$P(A) = \int_A f(x)dx$$

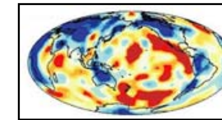
then the probability P(.) or the probability density f(.) represent the corresponding state of information on the parameter space (or sections of it).

Marginal probabilities: $$f_Y(y) = \int_X f(x,y)dx$$

… contains **all** information on parameter y. f(x,y) only contains information on the **correlation** (dependance) of x and y.

$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

States of information: perfect knowledge



The state of perfect knowledge:

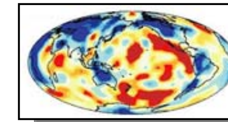
If we definitely know that the true value of x is $x=x_0$ the corresponding probability density is

$$f(\mathbf{x}) = \delta(\mathbf{x}-\mathbf{x}_0)$$

where $\delta(.)$ represents the Dirac delta function and $\int \delta(\mathbf{x}-\mathbf{x}_0) = 1$

This state is only useful in the sense that sometimes a parameter with respect to others is associated with much less error.

States of information: total ignorance



The state of total ignorance :

This is also termend the reference state of information (state of lowest information) called M(A) and the associated pdf is called the **non-informative** pdf $\mu(x)$

$$M(A) = \int_A \mu(\mathbf{x})d\mathbf{x}$$

where $\delta(.)$ represents the Dirac delta function and $\int \delta(\mathbf{x}-\mathbf{x}_0) = 1$

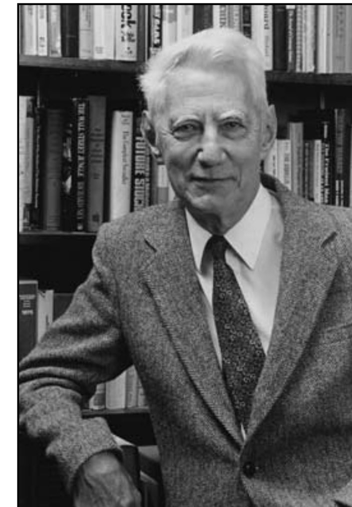
Example: Estimate the location of an event (party, earthquake, sunrise …) Does it make a difference whether we are in cartesian or in spherical coordinates?

$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

Shannon's information content

… Shannon must rank near the top of the list of the major figures of 20th century science …

Shannon invented the concept of quantifying the content of information (in a message, a formal system, etc.). His theory was the basis for digital Data transmission, data compression, etc. with enormous impact on today's daily things (CD, PC, digital phone, mobile phones, etc.)



Claude Shannon
1916-2001

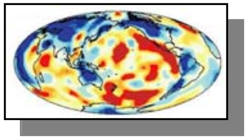
Definition: The information content for a discrete probabilistic system is

$$\boxed{H = \sum_i p_i \log p_i}$$

… but what does it really mean?

$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

Order, information, entropy



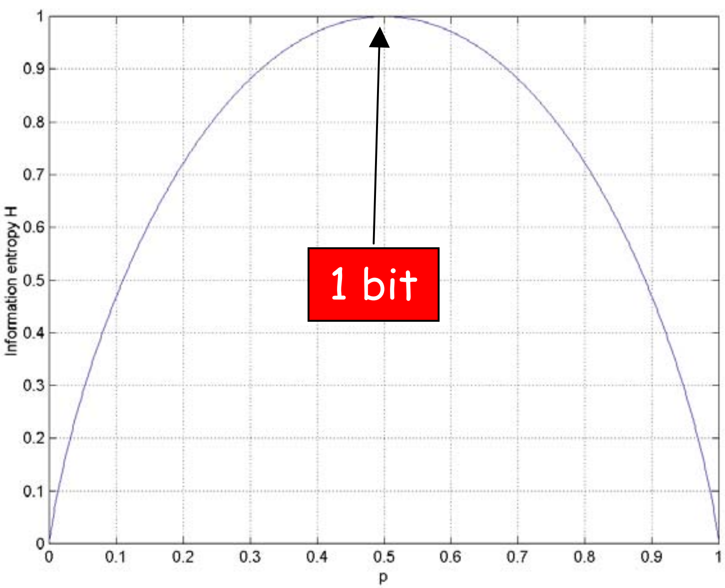
$$H = \sum_i p_i \log p_i$$

Let’s make a simple example:

The **information entropy** in the case of a of a system with two outcomes:

Event 1: p
Event 2: q=1-p

$$H = \sum_i p_i \log_2 p_i = -(p\log_2 p + q\log_2 q)$$



If we are certain of the outcome H=0.

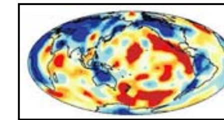
If uncertain, H is positive.

If all p_i are equal H has a maximum (most uncertainty, least order, maximum disorder)

<- this graph contains the definition of the most fundamental unit in information theory: guess!

$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

Bits, bytes, neps, digit



$$H = \sum_i p_i \log p_i$$

Shannon's formula is the basis for the units of information!

Log_2	->	Bits, Bytes
Log_e	->	Neps
Log_{10}	->	Digit

Some connections to physical entropy and disorder:

1111111111111111111 -> lots of order, no information, Shannon entropy small
0011010010110100010 -> low order, lots of information, Shannon entropy high

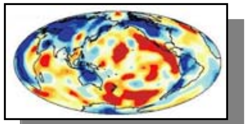
The first sequence can be expressed with one or two numbers, the second Sequence cannot be compressed.

In thermodynamics, entropy is a measure of microstates fileld in a crystal

Ice -> high order, small thermodynamic entropy, small Shannon entropy, not alot of information
Water -> disorder, large thermodynamic entropy, large Shannon entropy, wealth of information

$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

Shannon meets Tarantola



$$H = \sum_i p_i \log p_i$$

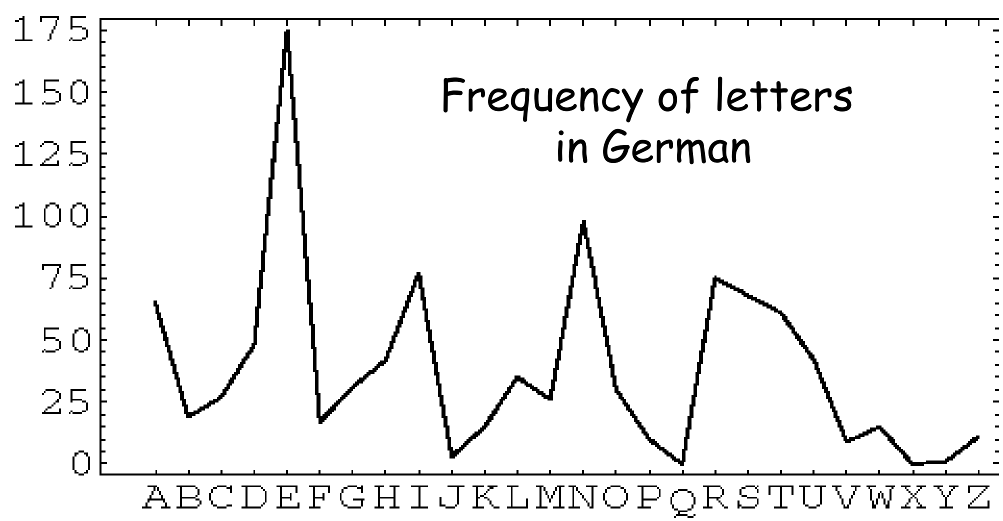
The generalization of Shannon's concept to the ideas of probabilistic inverse problems is

$$H(f,\mu) = \int f(\mathbf{x}) \log \frac{f(\mathbf{x})}{\mu(\mathbf{x})} d\mathbf{x}$$

… is called the information content of f(x). H(μ) represents the state of null information.

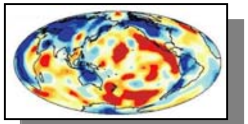
Finally: What is the information content of your name?

$$H = \sum_i p_i \log p_i$$



$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

Combining states of information



With basic principles from mathematical logic it can be shown that with two propositions f(x) (e.g. two data sets, two experiments, etc.) the combination of the two sources of information (with a logical **and**) comes down to

$$\boxed{\sigma(\mathbf{x}) = \frac{f_1(\mathbf{x})\, f_2(\mathbf{x})}{\mu(\mathbf{x})}}$$

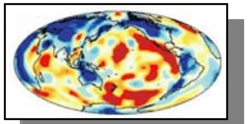
This is called the conjunction of states of information (Tarantola and Valette, 1982). Here μ(x) is the non-informative pdf and s(x) will turn out to be the **a posteriori** probability density function.

This equation is the basis for probabilistic inverse problems:

We will proceed to combine information obtained from measurements with information from a physical theory.

$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

Information from physical theories



Solving the forward problem is equivalent to predicting error free values of our data vector d, in the general case

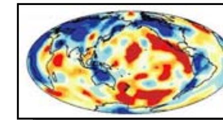
$$\boxed{d_{cal} = g(m)}$$

Examples:

- ground displacements for an earthquake source and a given earth model
- travel times for a regional or global earth model
- polarities and amplitudes for a given source radiation pattern
- magnetic polarities for a given plate tectonic model and field revearsal history
- shaking intensity map for a given earthquake and model
-....

But: Our modeling may contain errors, or may not be the right physical theory, How can we take this into account?

$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$



Information from physical theories

Following the previous ideas the most general way of describing information from a physical theory is by defining – for given values of model m - a probability density over the data space, i.e. a conditional probability density denoted by $\Theta(d|m)$.

Examples:

1. For an exact theory we have $\Theta(d \mid m) = \delta(d - g(m))$

2. Uncorrelated Gaussian errors

$$\boxed{\Theta(\mathbf{d} \mid \mathbf{m}) \propto \exp\left\{ -\frac{1}{2}(\mathbf{d}-\mathbf{g}(\mathbf{m}))^t C^{-1}(\mathbf{d}-\mathbf{g}(\mathbf{m})) \right\}}$$

where c is the covariance operator (a diagonal matrix) containing the variances.

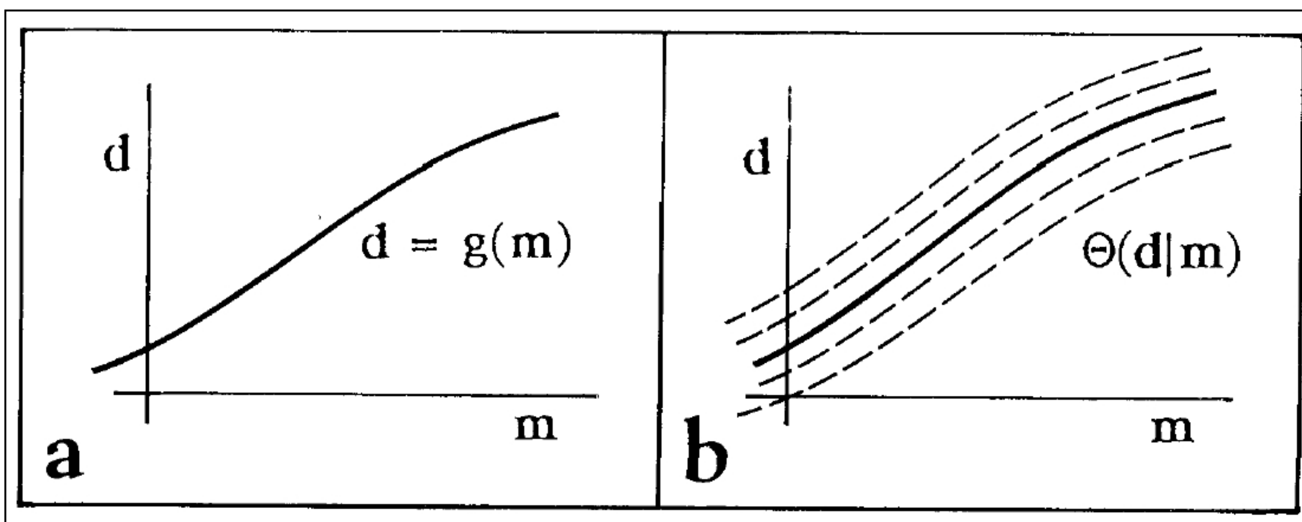
$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

Information from physical theories

$\Theta(d,m)$ summarized:

The expected correlations between model and data space can be described using the joint density function $\Theta(d,m)$. When there is an inexact physical theory (which is always the case), then the probability density for data d is given by $\Theta(d|m)\mu(m)$.

This may for example imply putting error bars about the predicted data d=g(m) … graphically

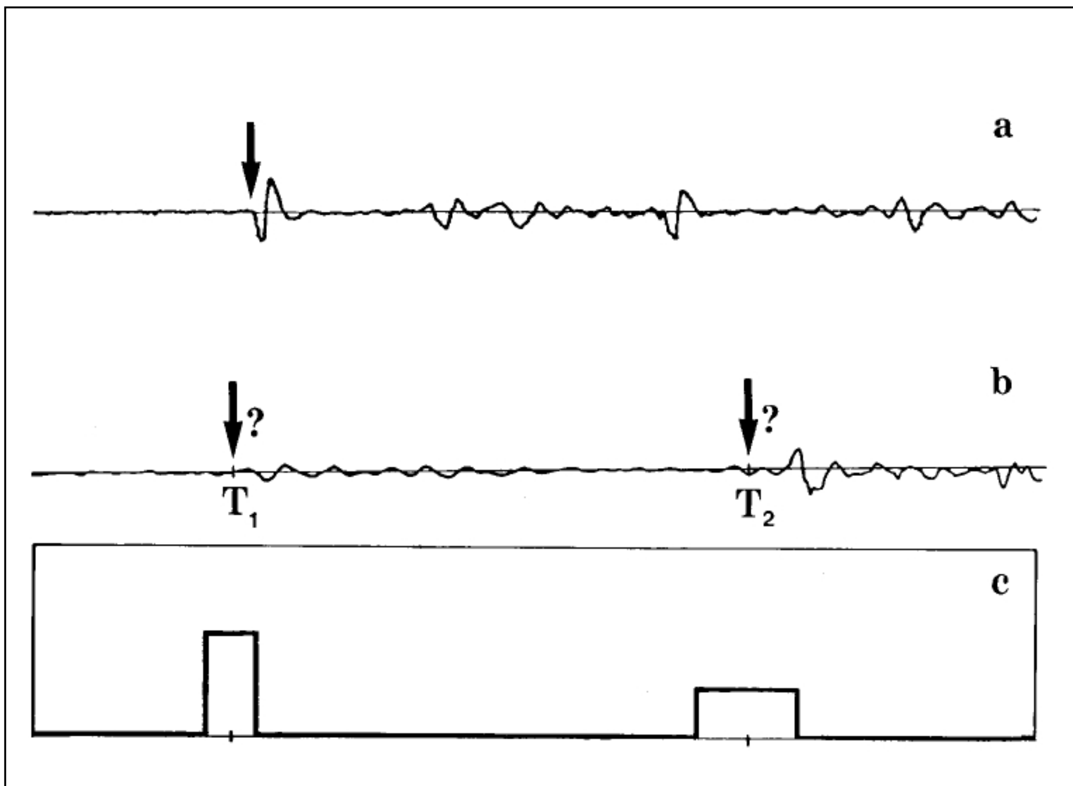


$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

Information from measurements

An experiment will give us **information on** the true values of observable parameters (but not actually the true values), we will call this pdf $\rho_D(d)$.

Example: Uncertainties of a travel time reading



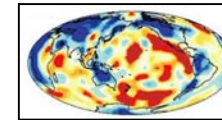
Good data

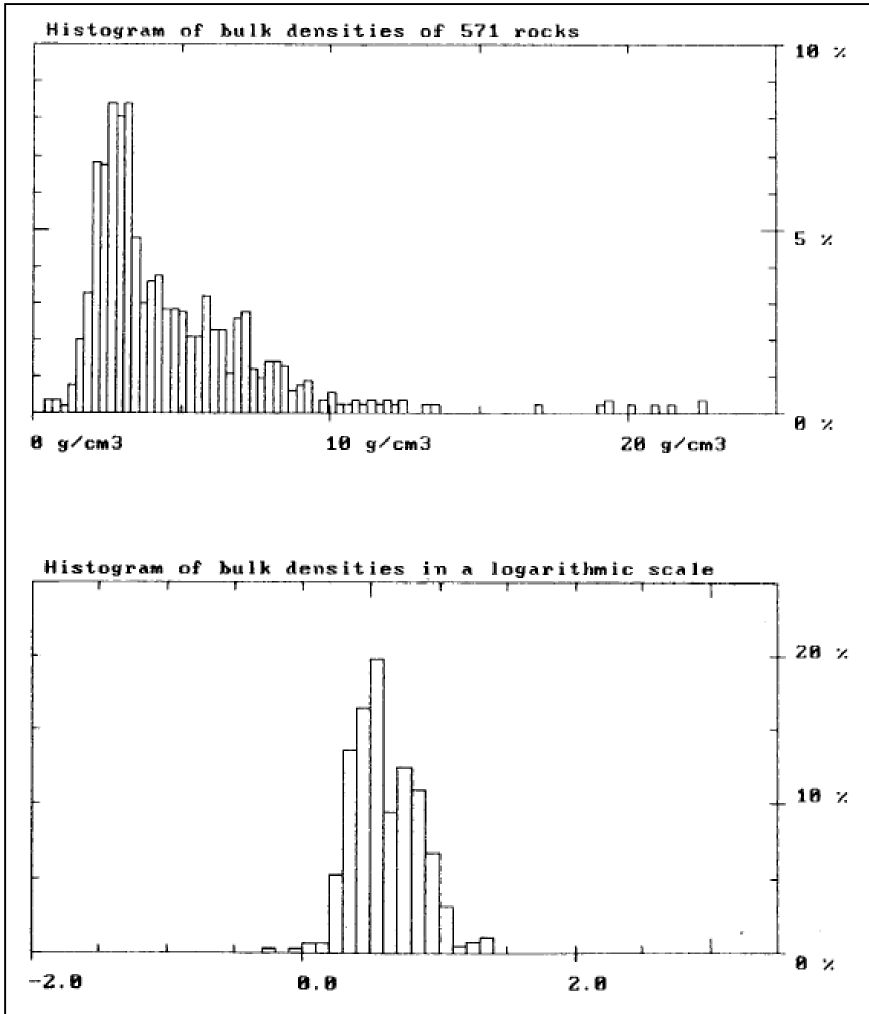
Noisy data

Uncertainty

$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

A priori information on model parameters





All the information obtained independently of the measurements on the model space is called **a priori information**. We describe this information using the pdf $\rho_M(m)$.

Example: We have no prior information $\rho_M(m)=\mu(m)$, where $\mu(m)$ is the non-informative prior.

Example: We are looking for a density model in the Earth (remember the treasure hunt). From sampling many many rocks we know what densities to expect in the Earth:

<- it looks like **lognormal** distributions are a Good way of describing some physical parameters

$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

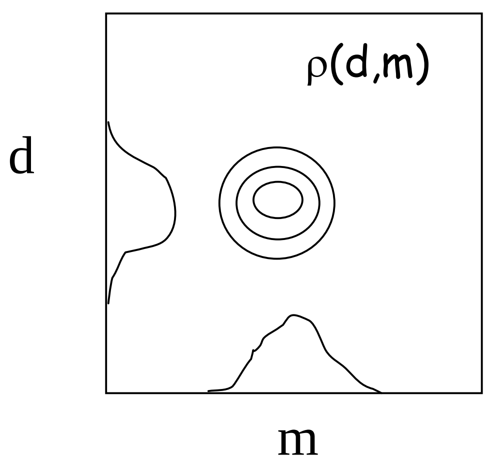
Joint prior information in model and data space

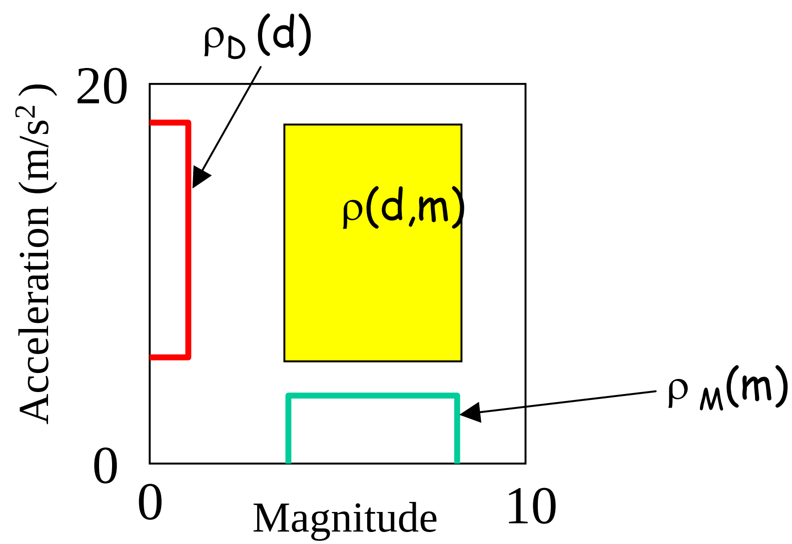
By definition, the a priori information on model parameters is independent of the a priori information on data parameters. We describe the information in the joint space DxM by

$\rho(d,m)= \rho_D(d)\, \rho_M(m)$

Example:

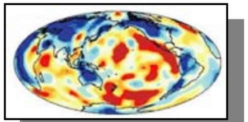
We observe the max. acceleration (data d) at a given site as a function of earthquake magnitude (model m). We expect earthquakes to have a magnitude smaller than 9 and larger than 4 (because the accelerometer would not trigger before). We also expect the max. acceleration not to exceed 18m/s and not be below 5 m/s^2 .





$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

The solution to the inverse problem



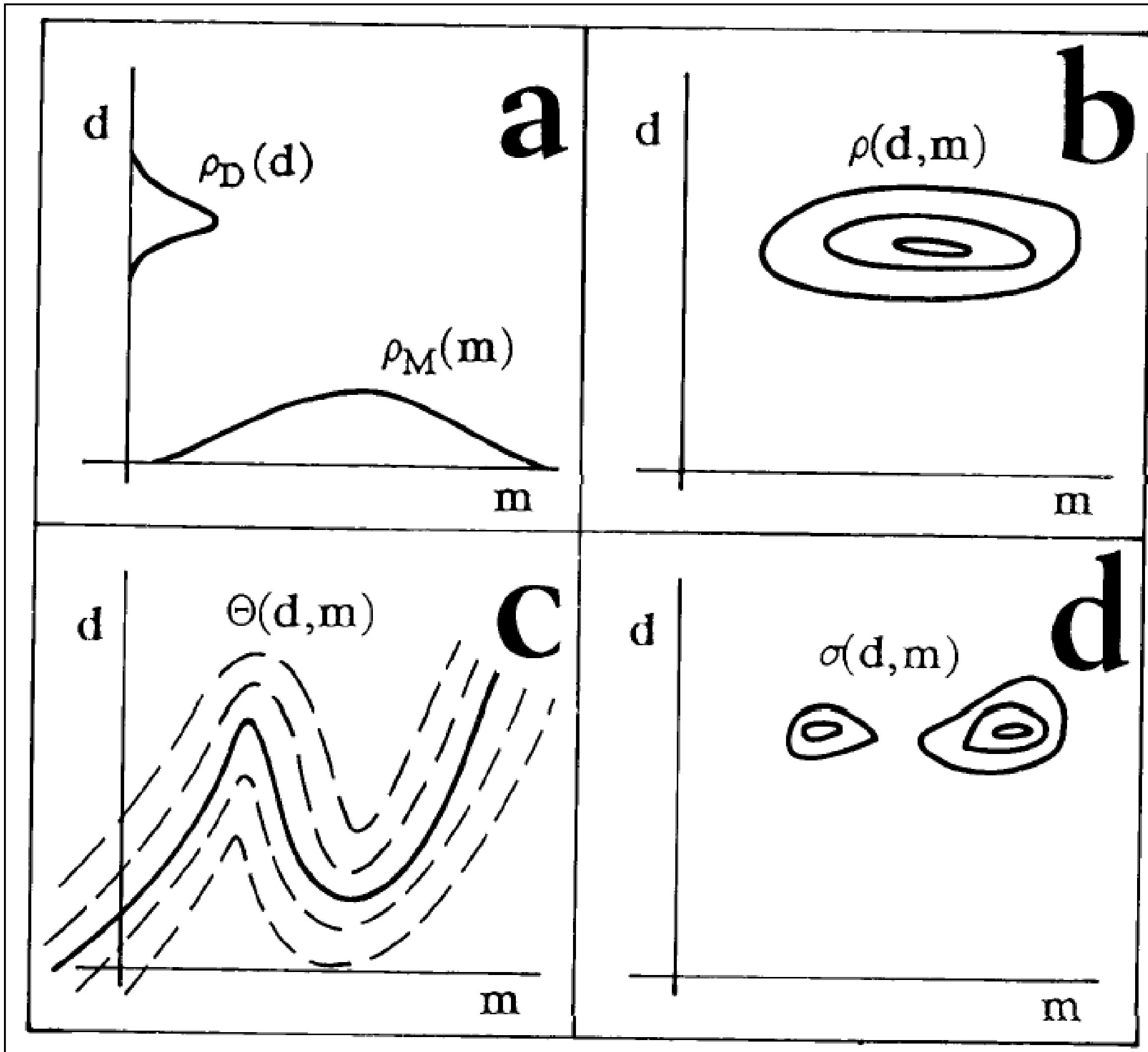
The information obtained **a priori** which we described with $\rho(d,m)$ is now combined with information from a physical theory which we decribe with $\Theta(d,m)$. Following the ideas of conjunction of states of information, we define the **a posteriori probability density funtion** as **the** solution to an inverse problem

$$\sigma(d,m) = \frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

Let's try and look at this graphically …

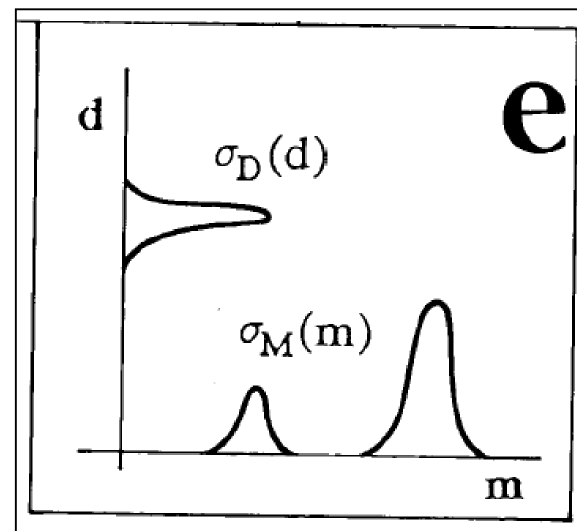
$$\sigma(d,m) = k\frac{\rho(d,m)\theta(d,m)}{\mu(d,m)}$$

The solution to the inverse problem

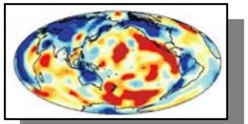


The (only) goal of this lecture is to understand these figures!

The rest is details …



Summary



Probability theory can be used to describe the **state of information** on a physical system. Actually it can be argued it is the only way of describing the necessarily subjective information we gain from physical experiments.

The key concept is to combine information which we know before the experiment (**a priori information**) with the information gained through observations and a physical theory.

The resulting **a posteriori probability density function** is the solution to the inverse problem.

The most difficult problem is how to obtain good samples of the A posteriori pdf, which will lead us to Monte Carlo methods, simulated Annealing and genetic algorithms.