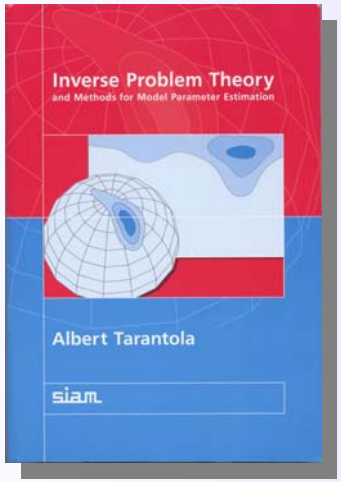
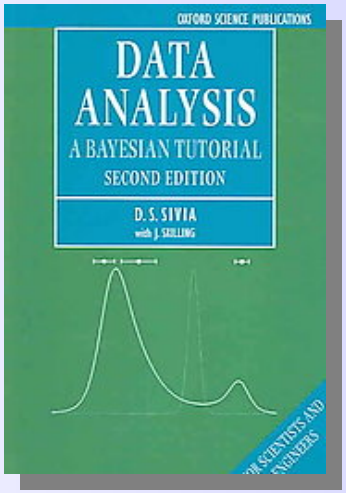
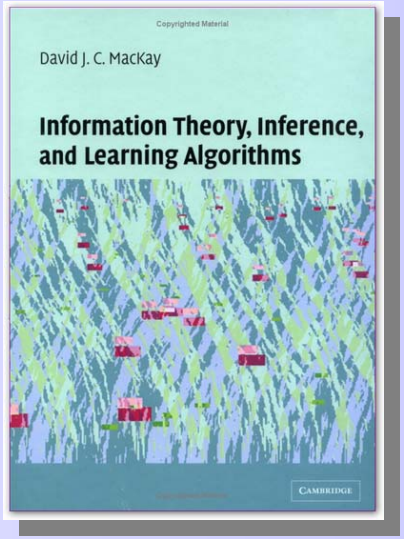
Probabilistic inference

Bayes theorem and all that....

Books

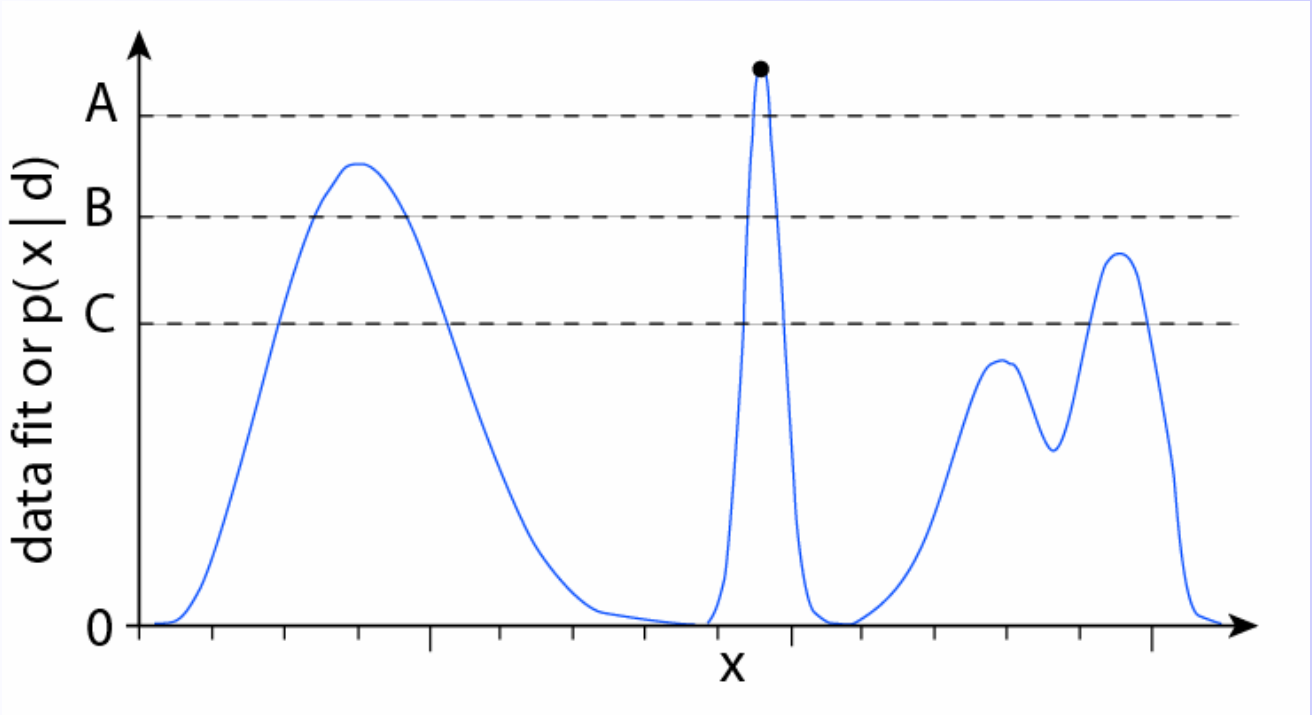

Inverse Problem Theory
and Methods for Model Parameter Estimation
Albert Tarantola
siam


OXFORD SCIENCE PUBLICATIONS
DATA ANALYSIS
A BAYESIAN TUTORIAL
SECOND EDITION
D. S. SIVIA
with J. SKILLING


Copyrighted Material
David J. C. MacKay
Information Theory, Inference, and Learning Algorithms
CAMBRIDGE

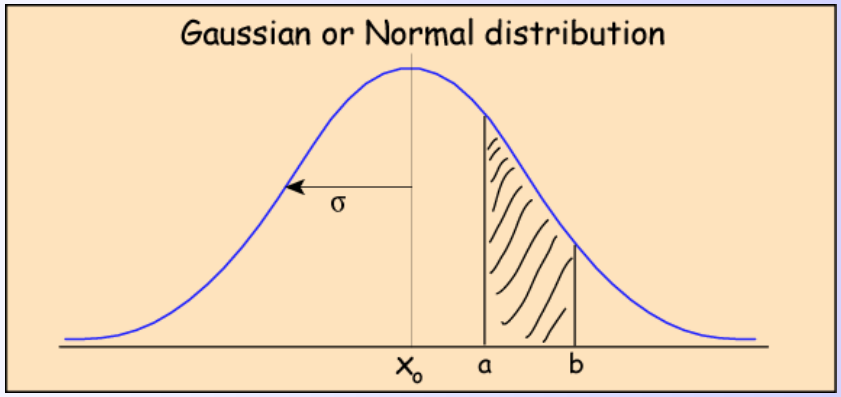
Highly nonlinear inverse problems

Multi-modal data misfit/objective function



What value is there in an optimal model ?

Probabilistic inference: History



$p(M|data, I)$

Mass of Saturn

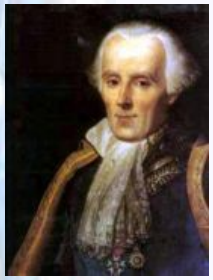
$$\int_{-\infty}^{\infty} p(x)dx = 1$$

(Laplace 1812)

$$Pr(x : a \leq x \leq b) = \int_a^b p(x)dx$$

We have already met the concept of using a probability density function p(x) to describe the state of a random variable.

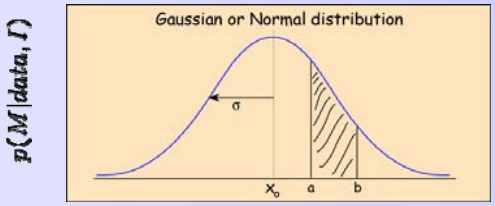
In the probabilistic (or *Bayesian*) approach, probabilities are also used to describe *inferences* (or *degrees of belief)* about x even if x itself is not a random variable.



Laplace (1812) rediscovered the work of Bayes (1763), and used it to constrain the mass of Saturn. In 150 years the estimate changed by only 0.63% !

But Laplace died in 1827 and then the arguments started...

Bayesian or Frequentist: the arguments



Mass of Saturn

Some thought that using probabilities to describe degrees of belief was too subjective and so they redefined probability as the *long run relative frequency* of a random event. This became the *Frequentist* approach.

To estimate the mass of Saturn the frequentist has to relate the mass to the data through a *statistic*. Since the data contain `random' noise probability theory can be applied to the statistic (which becomes the random variable !). This gave birth to the field of statistics !

But how to choose the statistic ?

`.. a plethora of tests and procedures without any clear underlying rationale'

(D. S. Sivia)

`Bayesian is subjective and requires too many guesses'

A. Frequentist

`Frequentist is subjective, but BI can solve problems more completely'

A. Bayesian

For a discussion see Sivia (2005, pp 8-11).

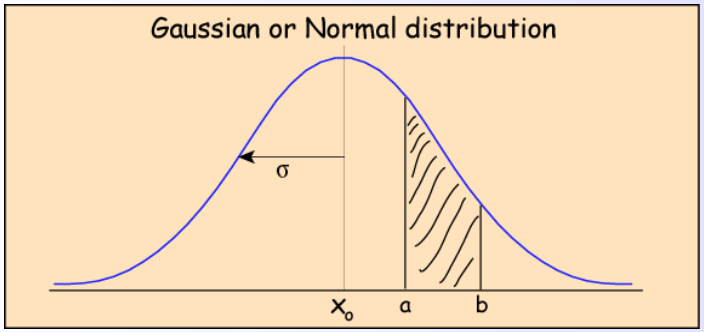
Probability theory: Joint probability density functions

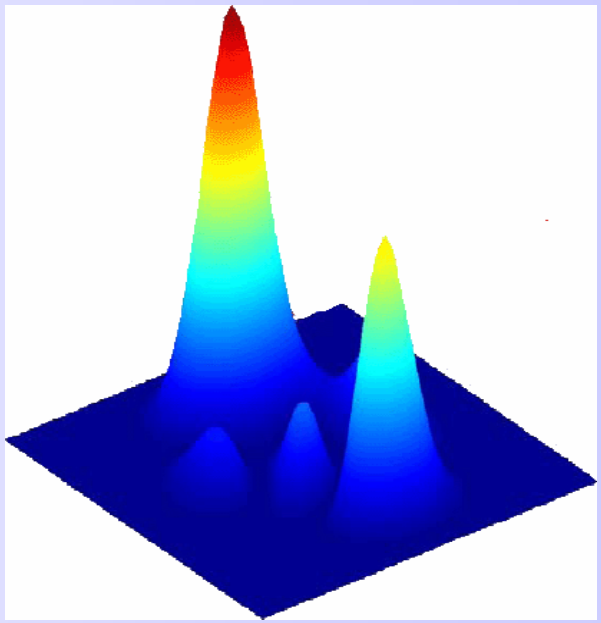
A PDF for variable x

$$p(x)$$

Joint PDF of x and y

$$p(x, y)$$





Probability is proportional to
area under the curve or surface

If x and y are independent their joint PDF is separable

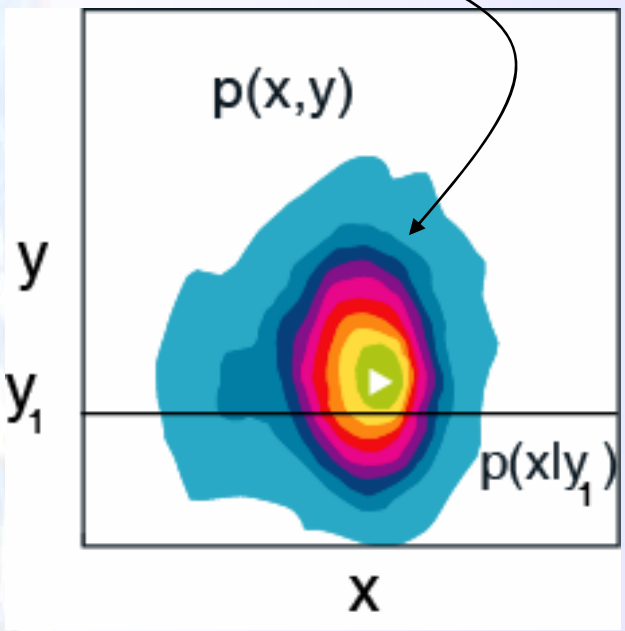
$$p(x, y) = p(x) \times p(y)$$

Probability theory: Conditional probability density functions

Joint PDF of x and y

"The PDF of x and y taken together"

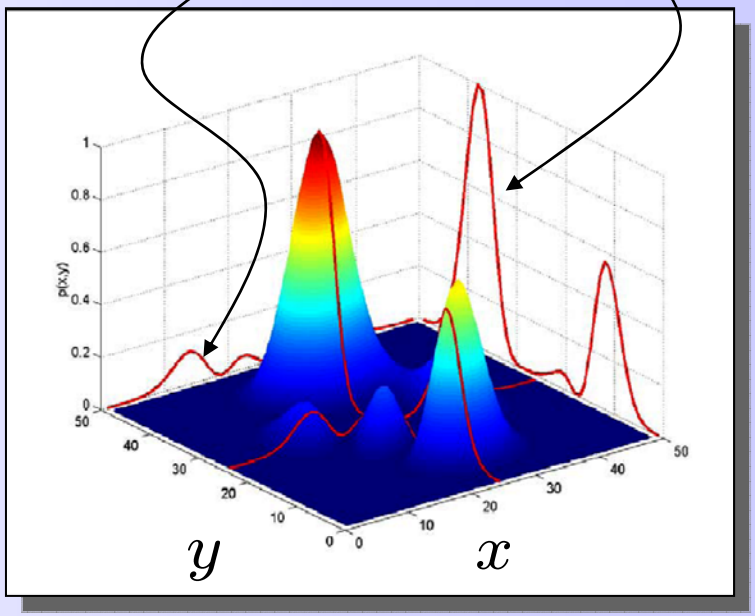
$p(x, y)$



Conditional PDFs

"The PDF of x given a value for y"

$p(x|y)$ $p(y|x)$



Relationship between joint and conditional PDFs

$$p(x, y) = p(x|y) \times p(y)$$

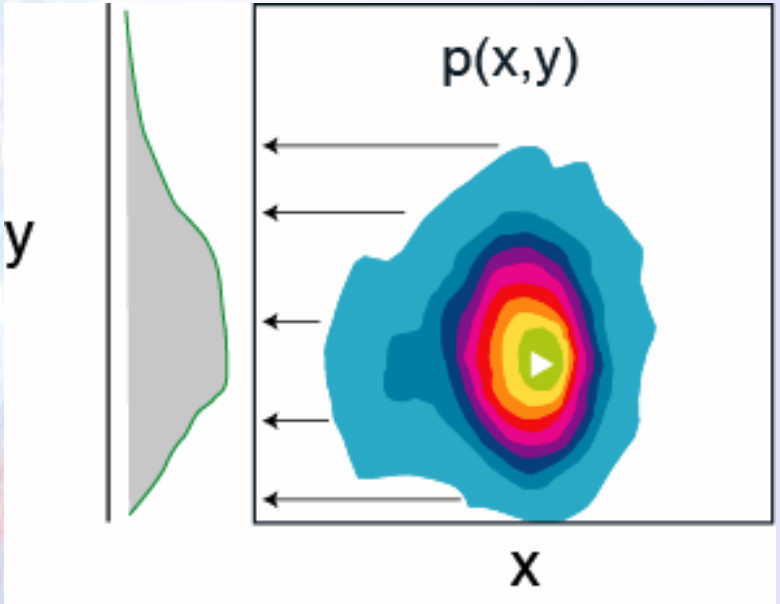
Probability theory: Marginal probability density functions

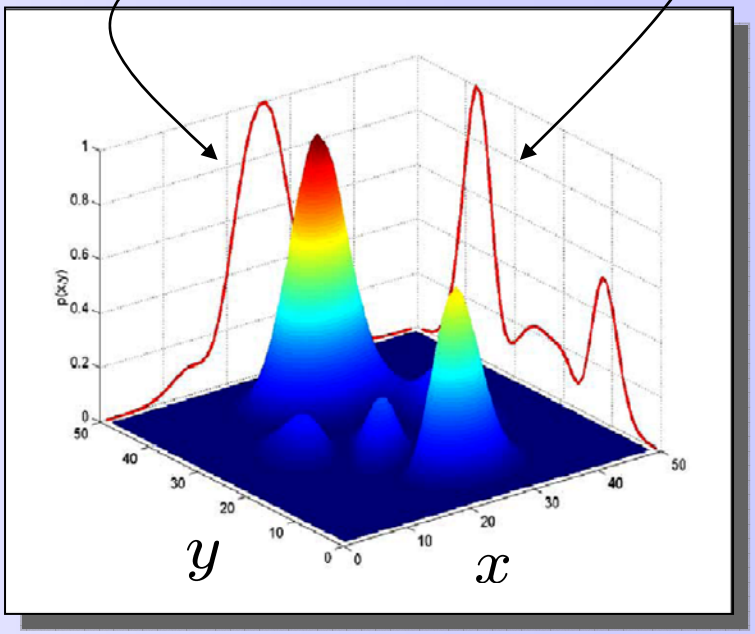
A marginal PDF is a summation of probabilities

Marginal PDFs

$$p(y) = \int p(x,y)dx$$

$$p(x) = \int p(x,y)dy$$



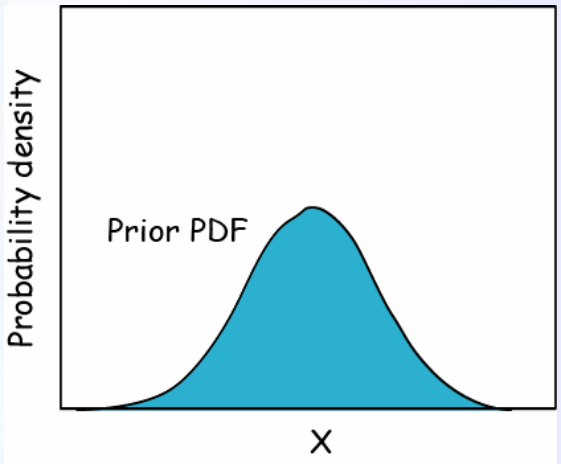


Relationship between joint, conditional and marginal PDFs

$$p(x,y) = p(x|y) \times p(y)$$

Prior probability density functions

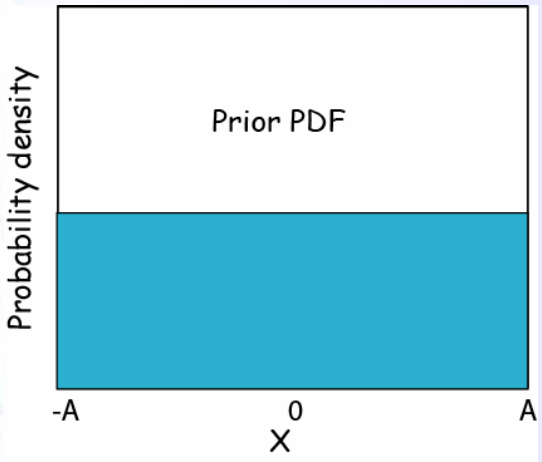
What we know from previous experiments, or what we guess...



$$p(x) = k \exp\left\{-\frac{(x - x_o)^2}{2\sigma^2}\right\}$$

$$p(\boldsymbol{m}) = k \exp\left\{-\frac{1}{2}(\boldsymbol{m} - \boldsymbol{m}_o)^T C_m^{-1}(\boldsymbol{m} - \boldsymbol{m}_o)\right\}$$

Beware: there is no such thing as a non-informative prior

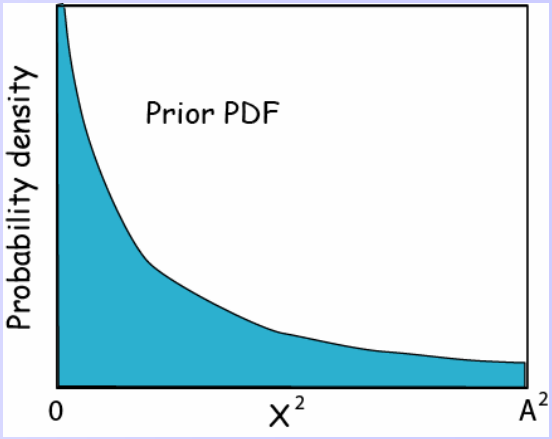


$$p(x)dx = p(y)dy$$

$$p(y) = p(x)\frac{dx}{dy}$$

$$p(x) = C$$

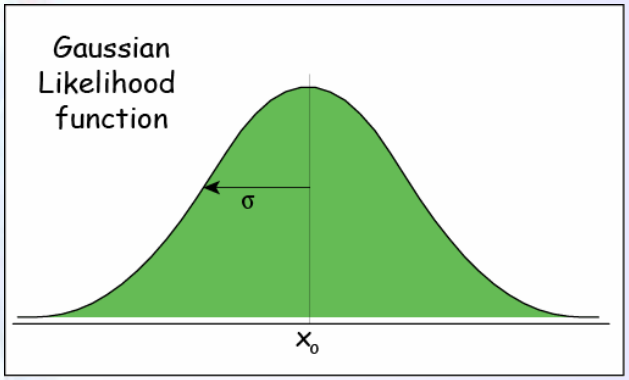
$$p(x^2) = \frac{C}{2x}$$



As A $\to \infty$ this is not proper !

Likelihood functions

The likelihood that the data would have occurred for a given model



$$p(d_i|x) = \exp\left\{-\frac{(x - x_{o,i})^2}{2\sigma_i^2}\right\}$$

$$p(\boldsymbol{d}|\boldsymbol{m}) = \exp\left\{-\frac{1}{2}(\boldsymbol{d} - G\boldsymbol{m})^T C_D^{-1}(\boldsymbol{d} - G\boldsymbol{m})\right\}$$

Maximizing likelihoods is what Frequentists do. It is what we did earlier.

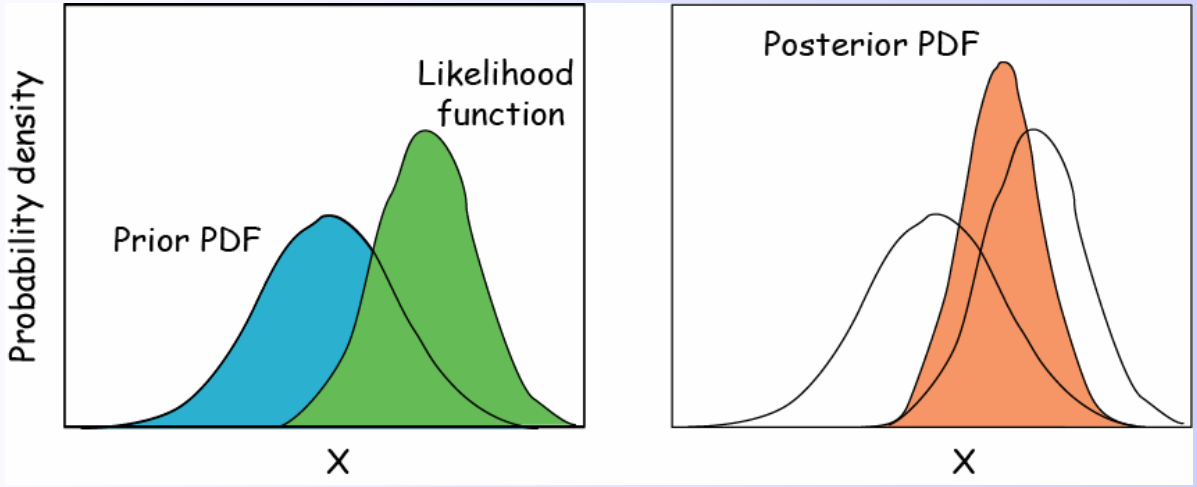
$$\max_{\boldsymbol{m}} \; p(\boldsymbol{d}|\boldsymbol{m}) = \min_{\boldsymbol{m}} -\ln(p(\boldsymbol{d}|\boldsymbol{m}))$$

$$= \min_{\boldsymbol{m}} \; (\boldsymbol{d} - G\boldsymbol{m})^T C_D^{-1}(\boldsymbol{d} - G\boldsymbol{m})$$

Maximizing the likelihood = minimizing the data prediction error

Bayes' theorem

All information is expressed in terms of probability density functions



Bayes' rule (1763)

Conditional PDFs — Assumptions

$$p(\boldsymbol{m}|\boldsymbol{d}, I) \propto p(\boldsymbol{d}|\boldsymbol{m}, I) \times p(\boldsymbol{m}|I)$$

data — model parameters

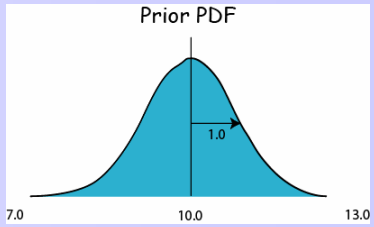
1702-1761

Posterior probability density $\propto$ Likelihood x Prior probability density

What is known after the data are collected — *Measuring fit to data* — *What is known before the data are collected*

Example: Measuring the mass of an object

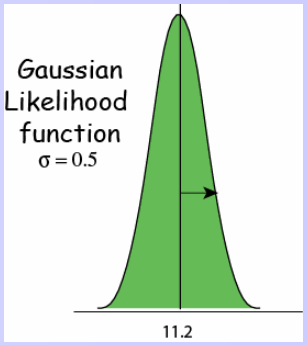
If we have an object whose mass, m, we which to determine. Before we collect any data we believe that its mass is approximately $10.0 \pm 1\mu g$. In probabilistic terms we could represent this as a Gaussian prior distribution



prior

$$p(m) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(m-10.0)^2}$$

Suppose a measurement is taken and a value 11.2 μ g is obtained, and the measuring device is believed to give Gaussian errors with mean 0 and σ = 0.5 μ g. Then the likelihood function can be written



Likelihood

$$p(d|m) = \frac{1}{0.5\sqrt{2\pi}} e^{-2(m-11.2)^2}$$

Posterior

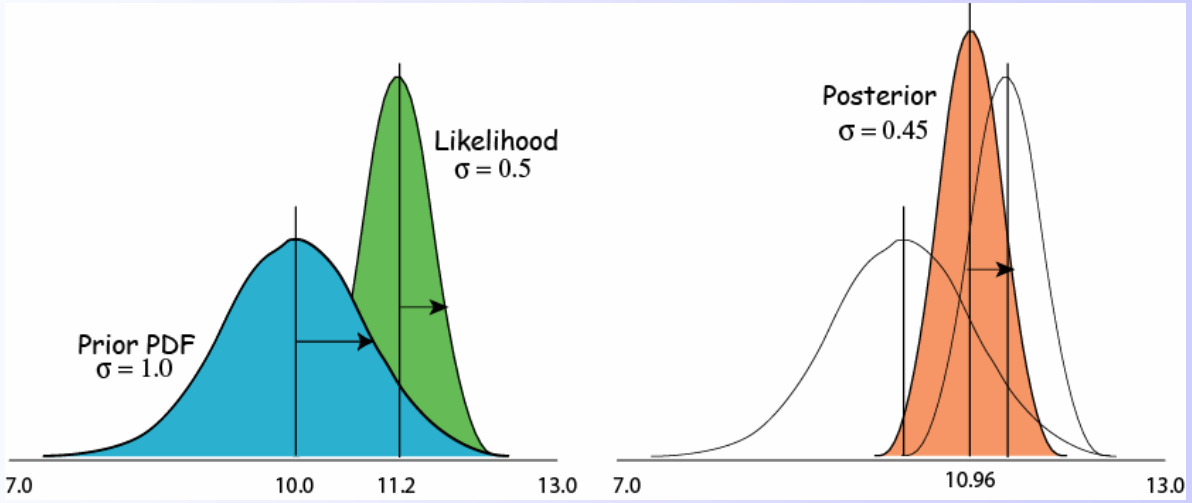
$$p(m|d) = \frac{1}{\pi} e^{-\frac{1}{2}(m-10.0)^2 - 2(m-11.2)^2}$$

$$p(m|d) \propto e^{\frac{-\frac{1}{2}(m-10.96)^2}{1/5}}$$

The posterior PDF becomes a Gaussian centred at the value of 10.96 μ g with standard deviation $\sigma = (1/5)^{1/2} \approx 0.45$.

Example: Measuring the mass of an object

The more accurate new data has changed the estimate of *m* and decreased its uncertainty



One data point problem

For the general linear inverse problem we would have

Prior: $$p(\boldsymbol{m}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{m}-\boldsymbol{m}_o)^T C_m^{-1}(\boldsymbol{m}-\boldsymbol{m}_o)\right\}$$

Likelihood: $$p(\boldsymbol{d}|\boldsymbol{m}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{d}-G\boldsymbol{m})^T C_d^{-1}(\boldsymbol{d}-G\boldsymbol{m})\right\}$$

Posterior PDF

$$\propto \exp\left\{-\frac{1}{2}[(\boldsymbol{d}-G\boldsymbol{m})^T C_d^{-1}(\boldsymbol{d}-G\boldsymbol{m}) + (\boldsymbol{m}-\boldsymbol{m}_o)^T C_m^{-1}(\boldsymbol{m}-\boldsymbol{m}_o)]\right\}$$

The biased coin problem

Suppose we have a suspicious coin and we want to know if it is biased or not ?

Let α be the probability that we get a head. $0 \leq \alpha \leq 1$

α = 1 : means we always get a head.

α = 0 : means we always get a tail.

α = 0.5 : means equal likelihood of head or tail.

We can collect data by tossing the coin many times

$$\{H, T, T, H, \ldots\}$$

We seek a probability density function for α given the data

$$p(\alpha|\boldsymbol{d}, I) \propto p(\boldsymbol{d}|\alpha, I) \times p(\alpha|I)$$

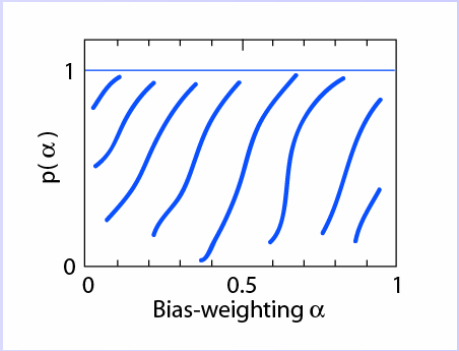
Posterior PDF $\propto$ Likelihood x Prior PDF

The biased coin problem

What is the prior PDF for α ?

Let us assume that it is uniform

$$p(\alpha|I) = 1, \quad 0 \leq \alpha \leq 1$$



What is the Likelihood function ?

The probability of observing R heads out of N coin tosses is

$$p(\boldsymbol{d}|\alpha, I) \propto \alpha^R (1-\alpha)^{N-R}$$

$$\boxed{p(\alpha|\boldsymbol{d}, I) \propto p(\boldsymbol{d}|\alpha, I) \times p(\alpha|I)}$$

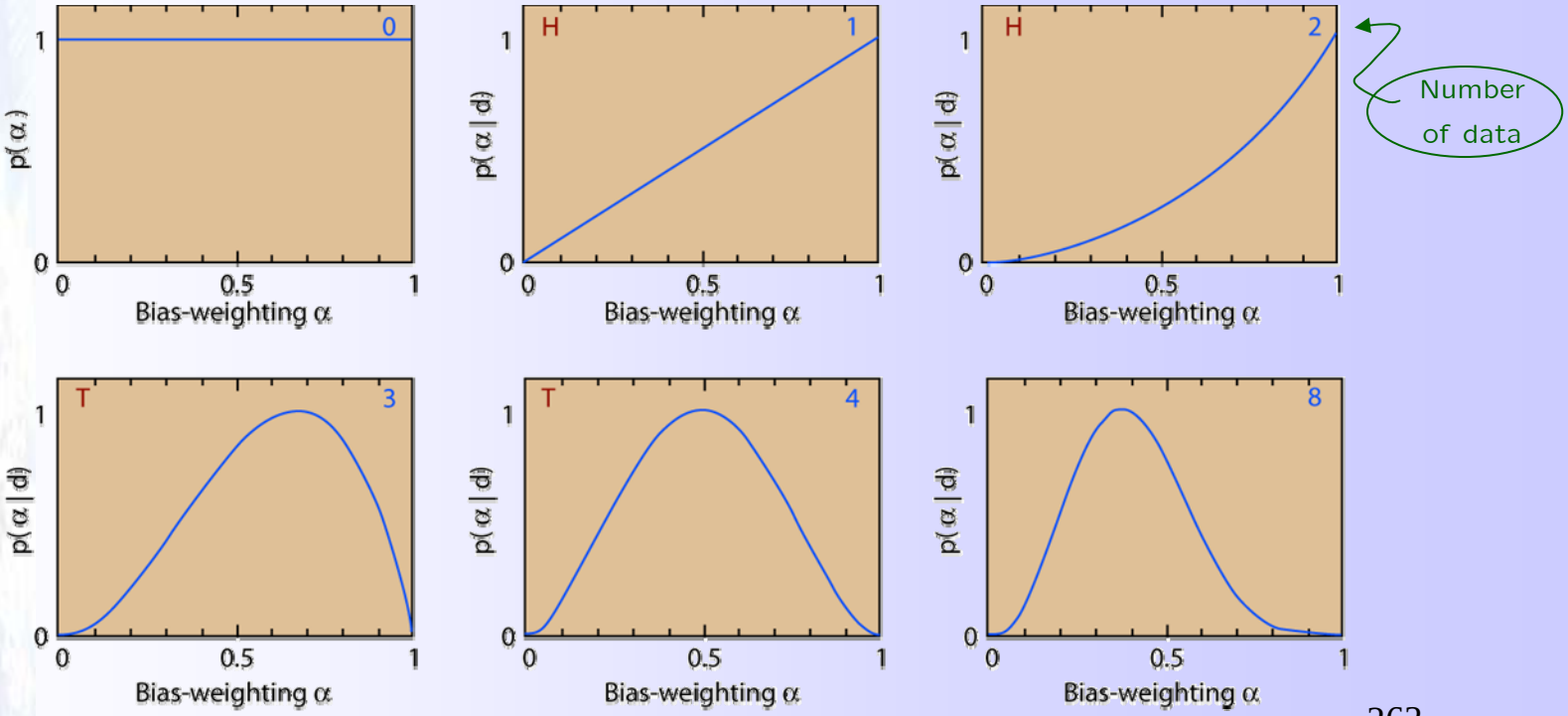
Posterior PDF $\propto$ Likelihood x Prior PDF

The biased coin problem

We have the posterior PDF for α given the data and our prior PDF

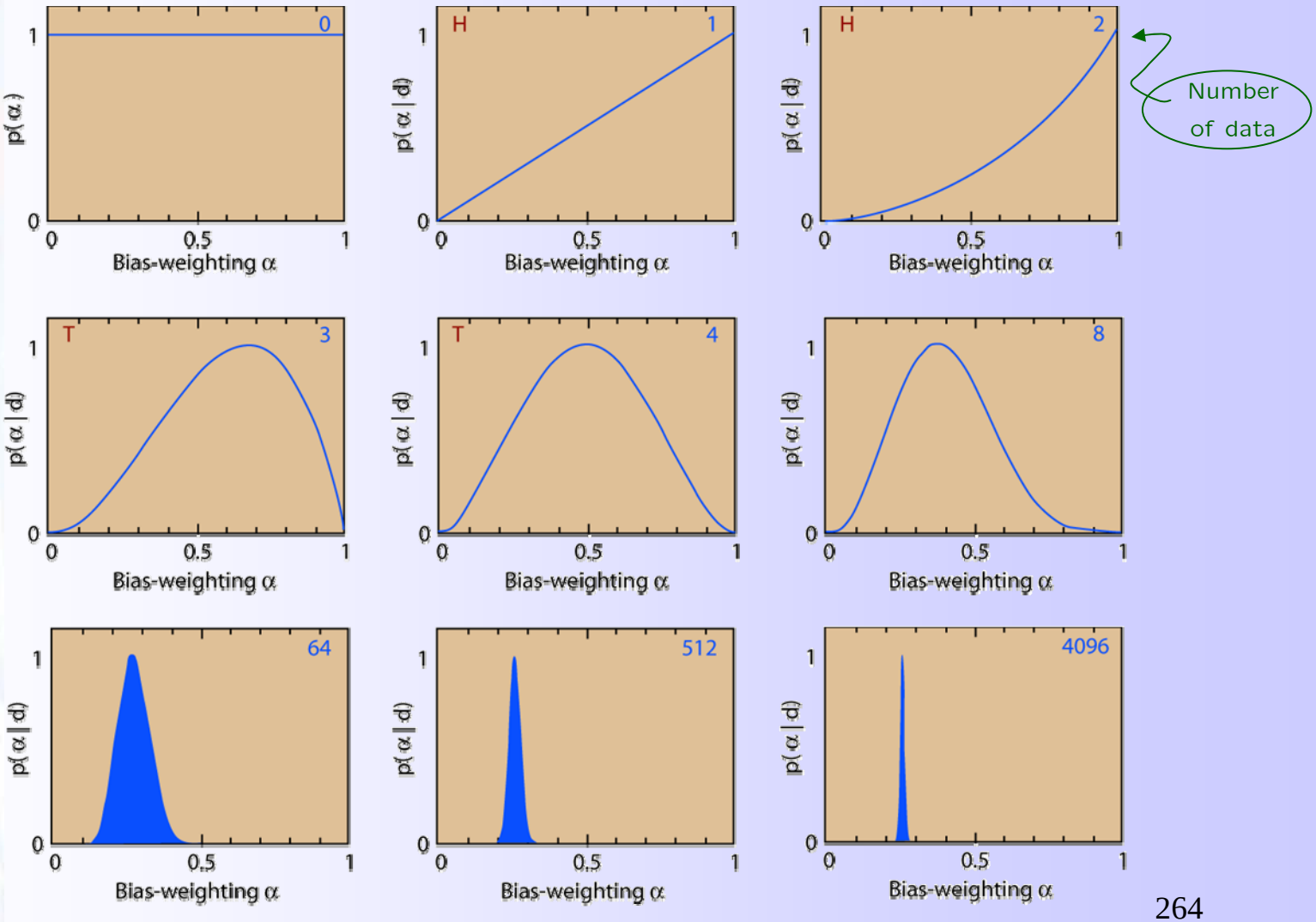
$$p(\alpha|\boldsymbol{d}, I) \propto \alpha^{R}(1-\alpha)^{N-R}$$

After N coin tosses let R = number of heads observed. Then we Can plot the probability density for p(α | d) as data are collected



The biased coin problem

$$p(\alpha|\boldsymbol{d}, I) \propto \alpha^{R}(1-\alpha)^{N-R}$$



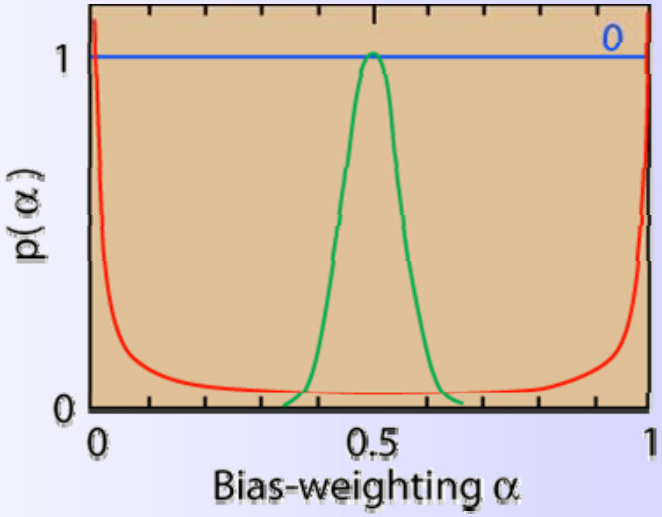
The biased coin problem

But what if three people had different opinions about the coin prior to collecting the data ?

Dr. Blue knows nothing about the coin.

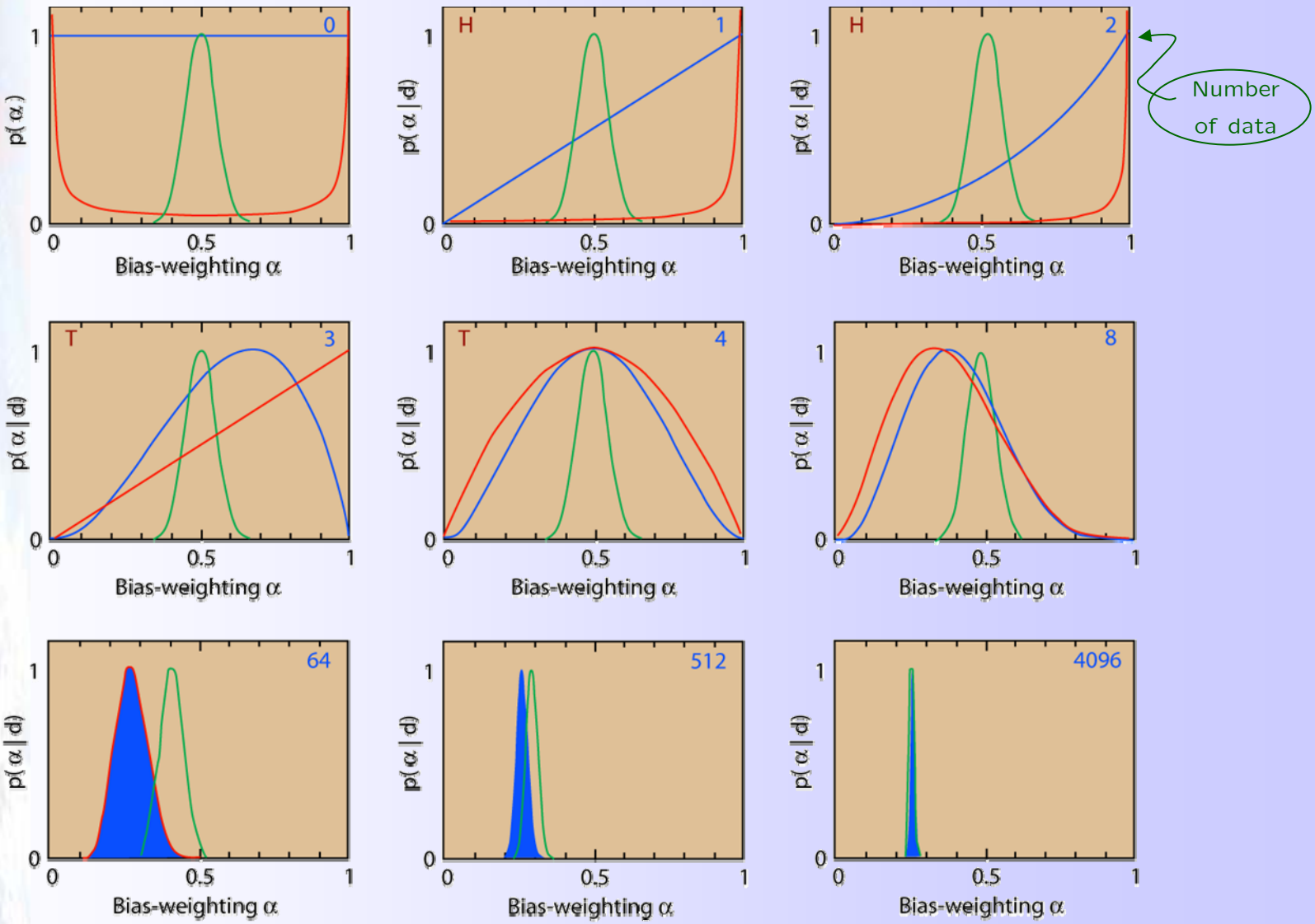
Dr. Green thinks the coin is likely to be almost fair.

Dr. Red thinks the coin is either highly biased to heads or tails.



$$p(\boldsymbol{d}|\alpha, I) \propto \alpha^{R}(1-\alpha)^{N-R}$$

The biased coin problem



The Lighthouse problem

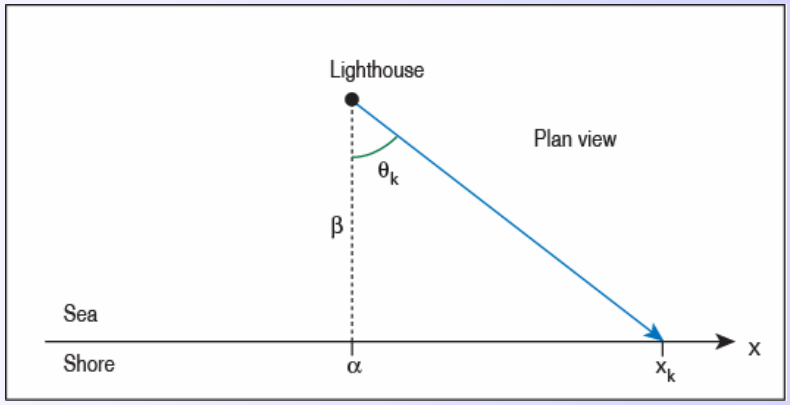


Photo-detectors along the shore record flashes from the light house but not the direction which the light came from.

Data: A total of *N* flashes have been recorded at positions x_k ($k=1,\ldots, N$)

Question: Where is the lighthouse ?

After Gull (1988)

How to choose a prior ?

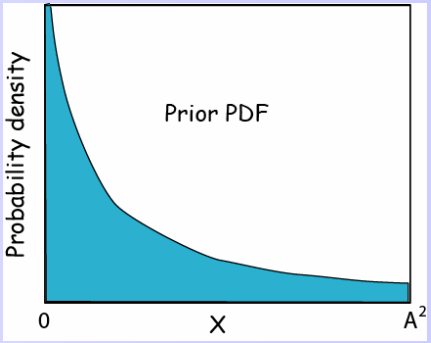
An often quoted weakness of Bayesian inversion is the subjectiveness of the prior.

If we know that x > 0 and that E{x} = μ.
What is an appropriate prior p(x) ?

$$H(X) = \int_{-\infty}^{\infty} p(x) \ln p(x) dx$$ Entropy

A solution is to choose the prior p(x) that maximizes entropy subject to satisfying the constraints. Using calculus of variations we get

$$p(x) = \frac{1}{\mu} e^{-x/\mu}, \quad x \geq 0$$

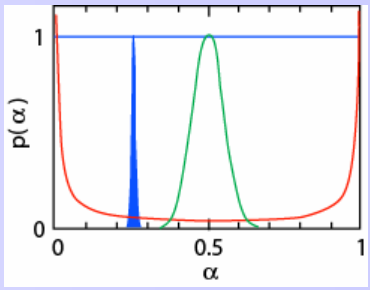


There is no such thing as a non-informative prior !

Recap: Probabilistic inference

- In the Bayesian treatment, all inferences are expressed in terms of probabilities.
- Bayes' theorem tells us how to combine a priori information with the information provided by the data, and all are expressed as PDFs.
- All Bayesian inference is relative. We always compare what we know after the data are collected to what we know before the data are collected. In practice this means comparing the a posteriori PDF with the a priori PDF.
- Bayesians argue that this is just a formalization of logical inference.
- Criticisms are that non-informative prior's do not exist, and hence we introduce information if prior's are assumed for convenience.
- The general framework is appealing but can not usually be applied when the number of unknowns is >= 10^3.

What can we do with the posterior PDF ?



- We could map it out over all of model space
  - Only feasible when dimension of the model space is small.
- We seek the maximum of posterior PDF (MAP) and its covariance

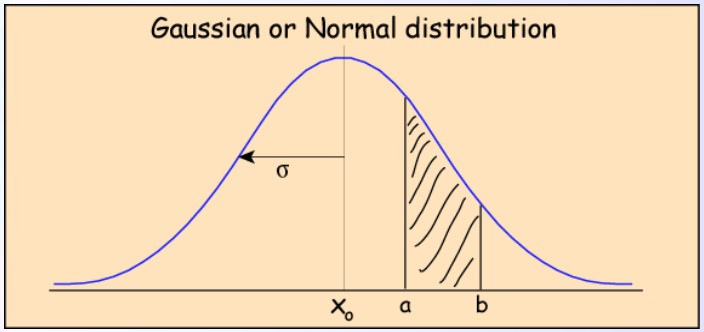
$$\phi(\boldsymbol{m},\boldsymbol{d}) = (\boldsymbol{d}-G\boldsymbol{m})^T C_d^{-1}(\boldsymbol{d}-G\boldsymbol{m})+(\boldsymbol{m}-\boldsymbol{m}_o)^T C_m^{-1}(\boldsymbol{m}-\boldsymbol{m}_o)$$

$$\max_{\boldsymbol{m}} \ \phi(\boldsymbol{m},\boldsymbol{d}); \quad \text{calculate } C_M$$

  - This is equivalent to the optimization approach earlier.
  - Would not make sense when the problem is multi-modal or when the covariance is not representative of its shape.
- We generate model (samples) whose density follows the posterior PDF. Posterior simulation is the main technique used in computational statistics.

In a Bayesian approach the complete Posterior PDF is the answer to the inverse problem, and we always look at its properties.

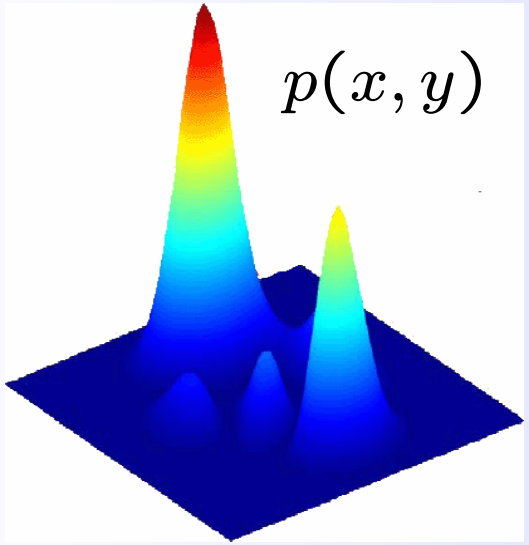
Sampling the posterior PDF



Histogram here

Probability is proportional to area under the curve or surface

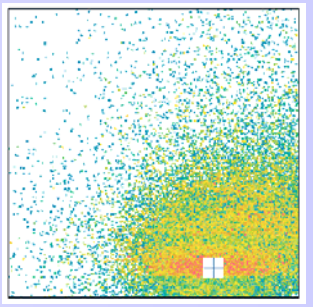
Sample density is proportional to area under the curve or surface



$$p(\boldsymbol{m}|\boldsymbol{d}, I) \propto p(\boldsymbol{d}|\boldsymbol{m}, I) \times p(\boldsymbol{m}|I)$$

Generating samples from the posterior PDF

There are several techniques for generating samples that follow general probability density functions.



- The transform method

$$y = f(x)$$

1-D $$p(y) = p(x)\left|\frac{dx}{dy}\right|$$

2-D $$p(y_1, y_2) = p(x_1, x_2)\left|\frac{\partial(x_1, x_2)}{\partial(y_1, y_2)}\right|$$

Jacobian determinant

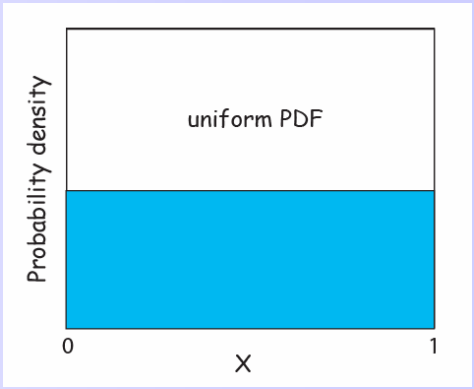
$$\left|\frac{\partial(x_1, x_2)}{\partial(y_1, y_2)}\right| = \left|\begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} \end{pmatrix}\right| = \left(\frac{\partial x_1}{\partial y_1}\right)\left(\frac{\partial x_2}{\partial y_2}\right) - \left(\frac{\partial x_2}{\partial y_1}\right)\left(\frac{\partial x_1}{\partial y_2}\right)$$

m-D $$p(y_1, y_2, \ldots, y_m) = p(x_1, x_2, \ldots, x_m)\left|\frac{\partial(x_1, x_2, \ldots, x_m)}{\partial(y_1, y_2, \ldots, y_m)}\right|$$

Example 1: a linear transform

Suppose we have a random variable x that is uniformly distributed between 0 and 1

$$p(x) = 1, \quad 0 \leq x \leq 1$$



Suppose we have the variable y where

$$y = (b-a)x + a$$

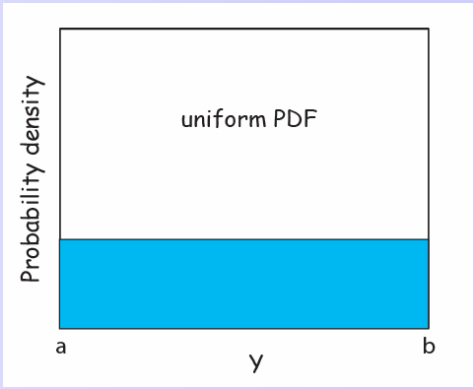
What is the PDF of y ?

$$x = \frac{y-a}{b-a}$$

Transformation rule gives

$$p(y) = p(x)\left|\frac{dx}{dy}\right|$$

$$\Rightarrow p(y) = \frac{1}{b-a}, \quad a \leq y \leq b$$



So y is a uniform random variable between a and b.

Example 2: a quadratic transform

Suppose we have x, such that

$$p(x) = \frac{1}{2A}, \quad -A \leq x \leq A$$

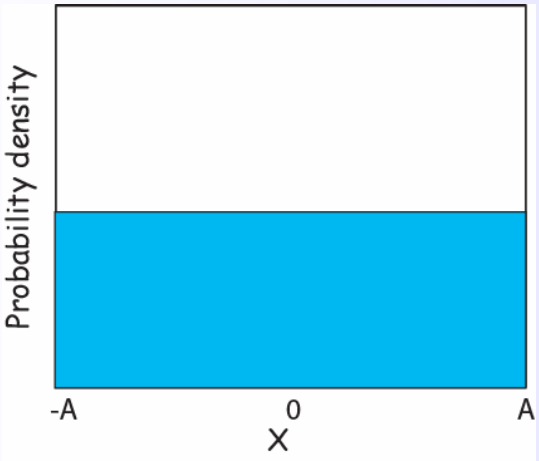
Suppose we have y such that

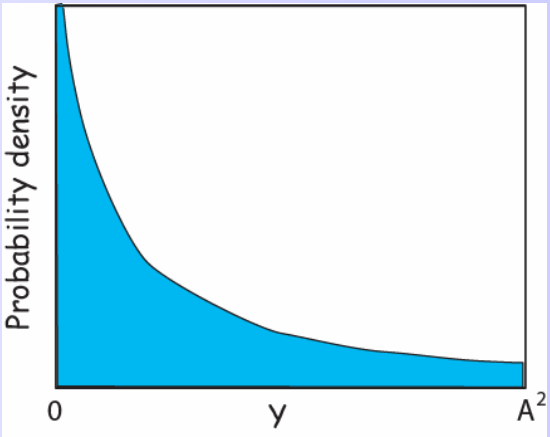
$$y = x^2$$

What is the PDF of y ?

$$p(y) = p(x)\left|\frac{dx}{dy}\right|$$

$$p(y) = \frac{1}{4Ay^{1/2}}$$





Example 3: Box-Muller transform

Suppose we have two independent uniform random variables, x_1 and x_2

$$p(x_1) = p(x_2) = 1, \quad 0 \le x_1, x_2 \le 1$$

Consider the transform to y_1, y_2

$$y_1 = \sqrt{-2 \ln x_1} \cos 2\pi x_2, \qquad y_2 = \sqrt{-2 \ln x_1} \sin 2\pi x_2$$

Inverting we get

$$x_1 = \exp\left[-\frac{1}{2}(y_1^2 + y_2^2)\right] \qquad x_2 = \frac{1}{2\pi} \tan^{-1} \frac{y_2}{y_1}$$

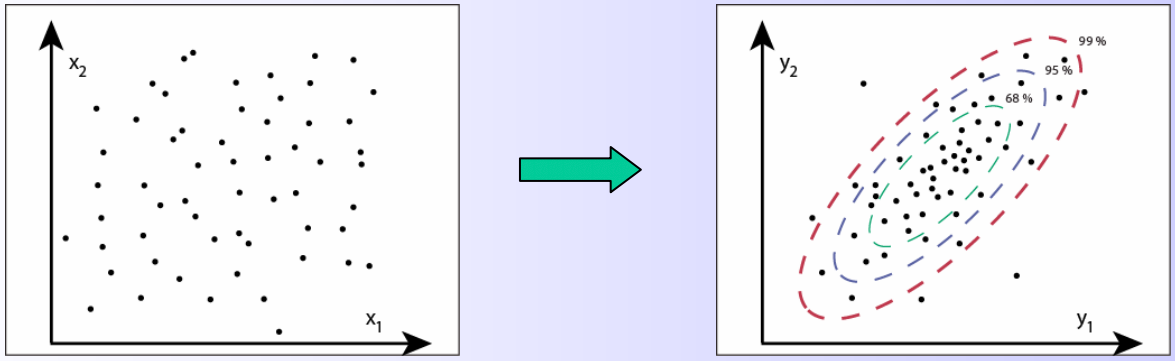
$$p(y_1, y_2) = p(x_1, x_2) \left| \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} \right|$$

$$\left| \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} \right| = - \left[\frac{1}{\sqrt{2\pi}} e^{-y_1^2/2} \right] \left[\frac{1}{\sqrt{2\pi}} e^{-y_2^2/2} \right]$$

Example 3: Box-Muller transform

So y_1 and y_2 have independent Gaussian distributions

$$p(y_1, y_2) = -\left[\frac{1}{\sqrt{2\pi}} e^{-y_1^2/2}\right]\left[\frac{1}{\sqrt{2\pi}} e^{-y_2^2/2}\right]$$

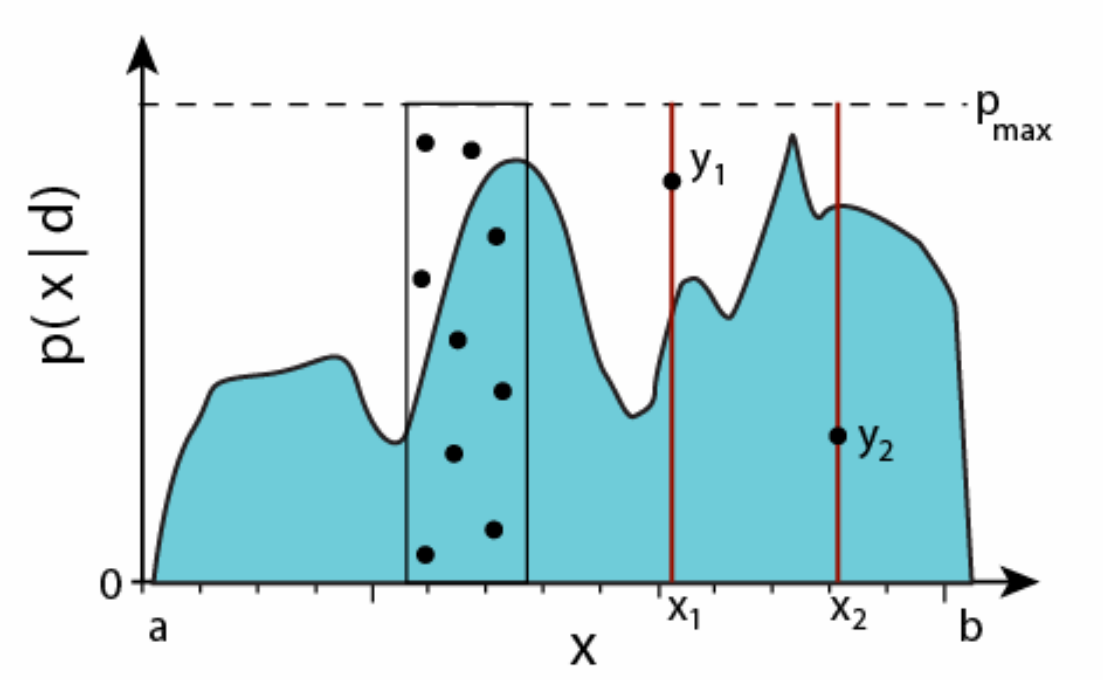


Transform methods are efficient because we get one set of output variables (y_1 and y_2) for each set of input, (x_1, x_2) variables.

But transformations can not always be inverted. The transform method will be limited to those cases for which an invertible transform can be found. General PDFs $p(x_1, x_2, \ldots)$ can not be handled.

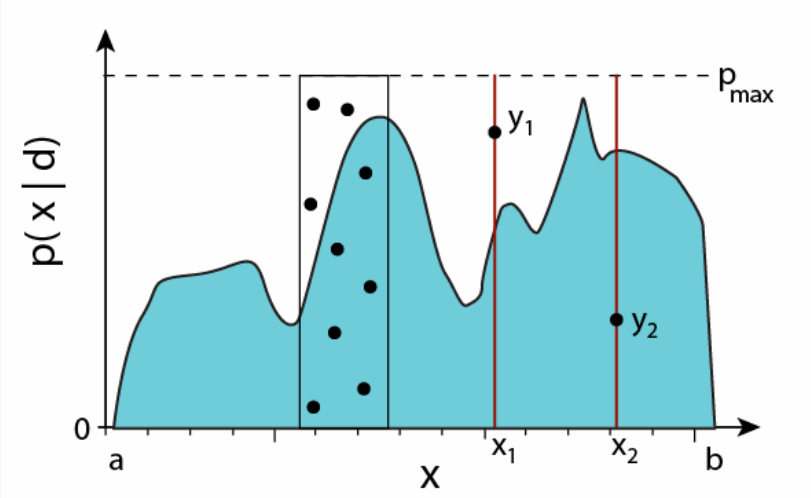
Generating samples from an arbitrary posterior PDF

- The rejection method



Generating samples from the posterior PDF

- Rejection method



But this requires us to know P_{max}

Step 1: generate a uniform random variable, x_i between a and b

$$p(x_i) = \frac{1}{(b-a)}, \quad a \leq x_i \leq b$$

Step 2: generate a second uniform random variable, y_i

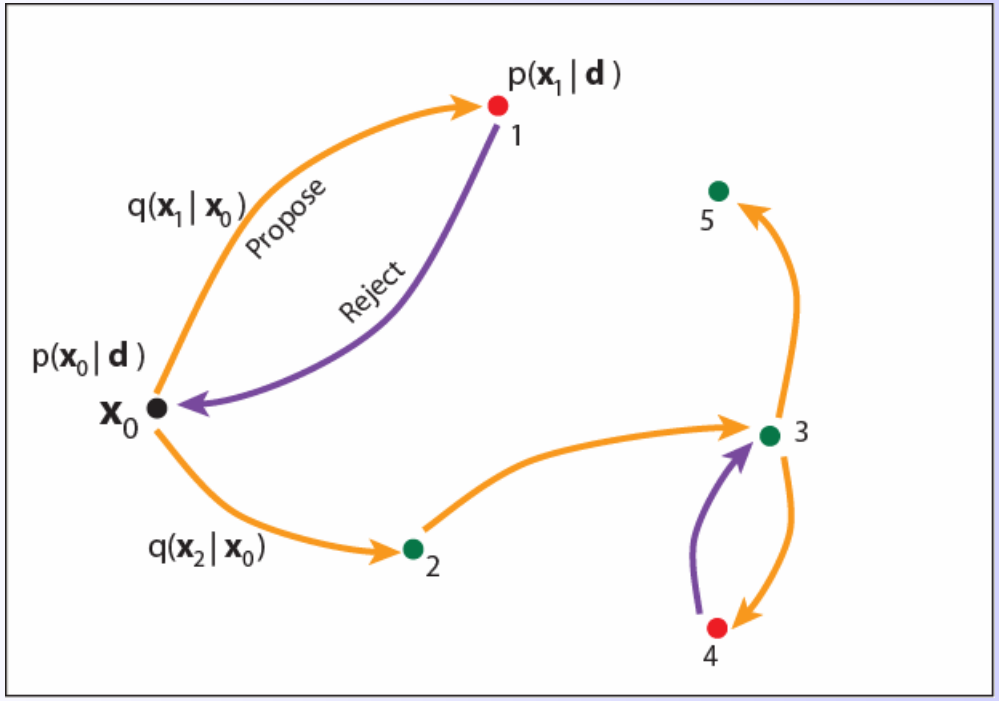
$$p(y_i) = \frac{1}{p_{max}}, \quad 0 \leq y_i \leq p_{max}$$

Step 3: accept x_i if $y_i \leq p(x_i|d)$ otherwise reject

Step 4: go to step 1

Generating samples from the posterior PDF

- Markov chain Monte Carlo (MCMC)



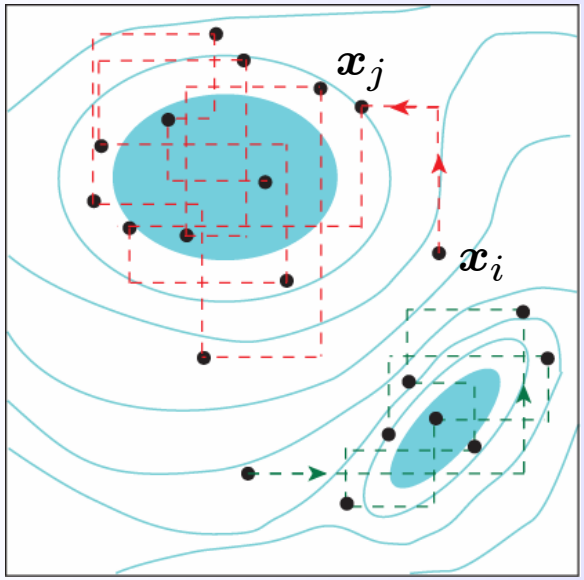
Metropolis algorithm random walk

Step 1: From point x_i generate a candidate model x_j, using a probabilistic *proposal density*

Step 2: Accept the new point with a probability dependent on the posterior PDF at that the new point.

Generating samples from the posterior PDF

- Markov chain Monte Carlo (MCMC)



Importance sampling

Metropolis algorithm
random walk

$$x_i \rightarrow x_j$$

Step 1: From point x_i generate a candidate model x_j, using a *proposal density*

$$q(x_j|x_i)$$

Probabilistic model generation

$$q(x_j|x_i) = q(x_i|x_j)$$

Step 2: Accept the new point randomly with probability p_a

Probabilistic acceptance criterion

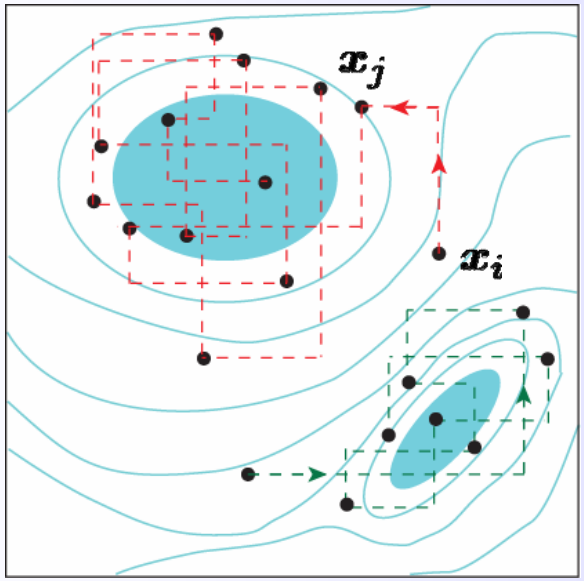
$$p = \min\left[1, \frac{q(x_i|x_j) \times p(x_i|d)}{q(x_j|x_i) \times p(x_j|d)}\right] = \min\left[1, \frac{p(x_i|d)}{p(x_j|d)}\right]$$

$$\Rightarrow p = \begin{cases} \frac{p(x_i|d)}{p(x_j|d)} & : \quad p(x_j|d) \geq p(x_i|d) \\ 1 & : \quad \text{otherwise} \end{cases}$$

If the step is rejected then we stay at x_i and return to step 1

Generating samples from the posterior PDF

Markov chain Monte Carlo (MCMC)



Importance sampling

Acceptance probability

$$p = \left[1, \frac{p(\boldsymbol{x}_i|\boldsymbol{d})}{p(\boldsymbol{x}_j|\boldsymbol{d})}\right]$$

Under general conditions it can be proven that this algorithm converges to the desired posterior distribution $p(\boldsymbol{x}|\boldsymbol{d})$ asymptotically.

- This works for general distributions and only requires evaluation of the posterior PDF for any input model, x.
- Efficiency depends on the choice of proposal distribution $q(\boldsymbol{x}_j|\boldsymbol{x}_i)$ and the chain may require `thinning' to remove dependence between samples.

This is the workhorse technique in Bayesian statistics

Example: MCMC on a 1-D PDF

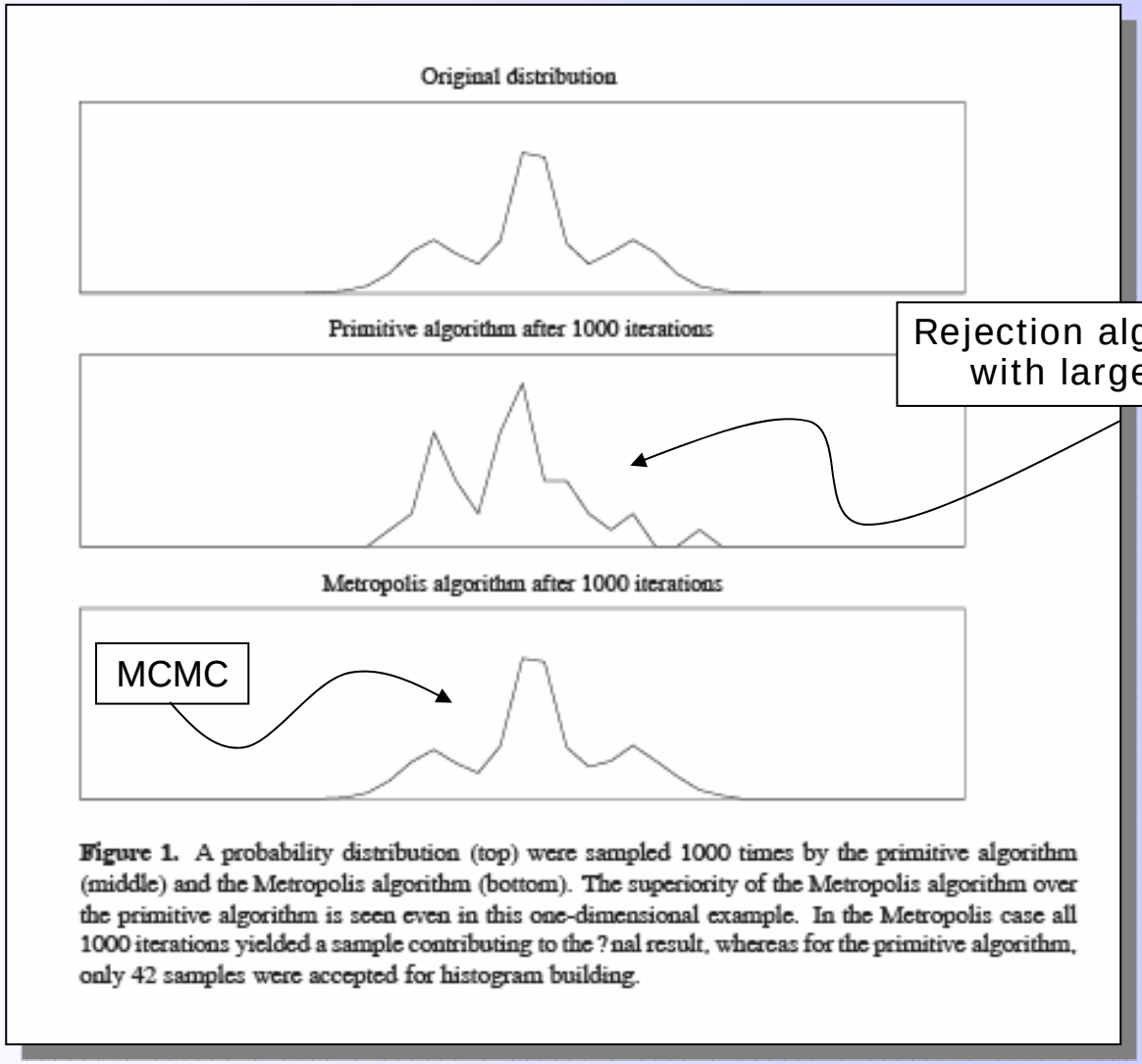


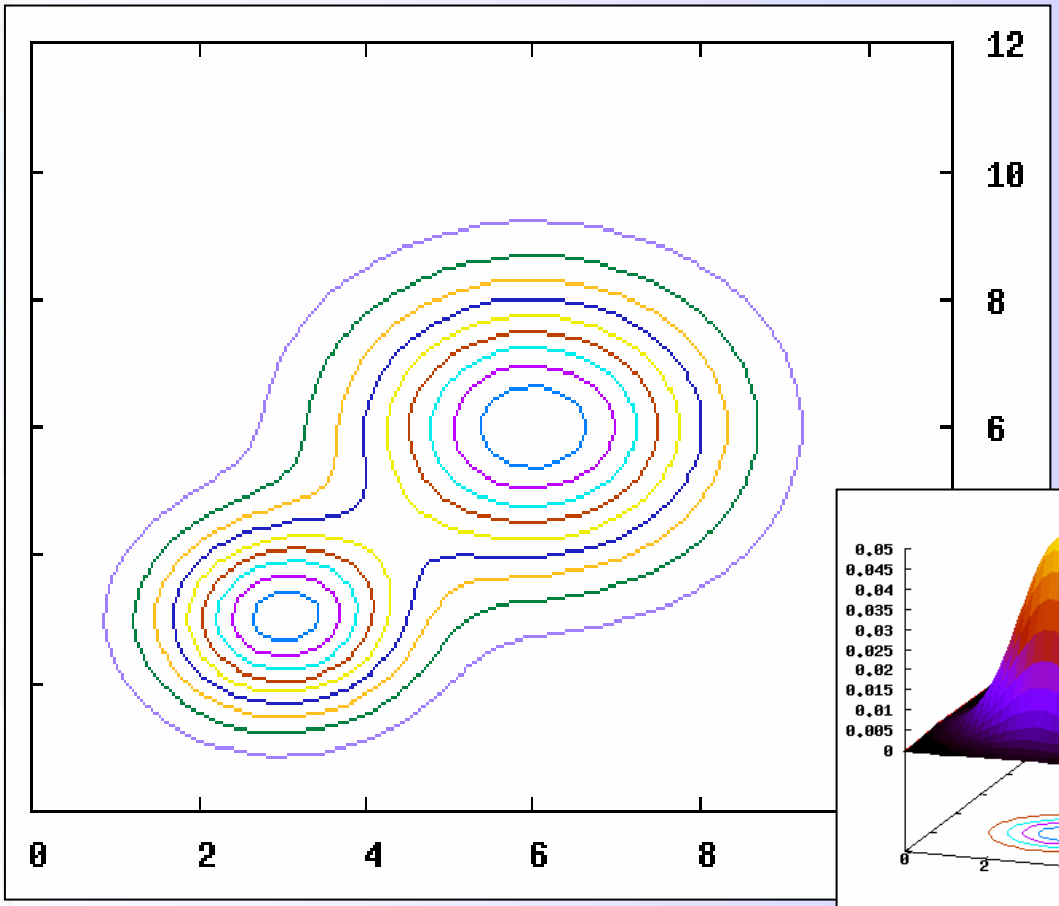
Figure 1. A probability distribution (top) were sampled 1000 times by the primitive algorithm (middle) and the Metropolis algorithm (bottom). The superiority of the Metropolis algorithm over the primitive algorithm is seen even in this one-dimensional example. In the Metropolis case all 1000 iterations yielded a sample contributing to the ?nal result, whereas for the primitive algorithm, only 42 samples were accepted for histogram building.

From Mosegaard and Sambrdge 2002

Example: MCMC on a multi-modal PDF

A sum of two 2-D Gaussian distributions

$$p(x,y) = \frac{1-\lambda}{2\pi\sigma_{x_1}\sigma_{y_1}} \exp - \left\{ \frac{(x-x_1^o)^2}{2\sigma_{x_1}^2} + \frac{(y-y_1^o)^2}{2\sigma_{y_1}^2} \right\} + \frac{\lambda}{2\pi\sigma_{y_2}\sigma_{y_2}} \exp - \left\{ \frac{(x-x_2^o)^2}{2\sigma_{x_2}^2} + \frac{(y-y_2^o)^2}{2\sigma_{y_2}^2} \right\}$$



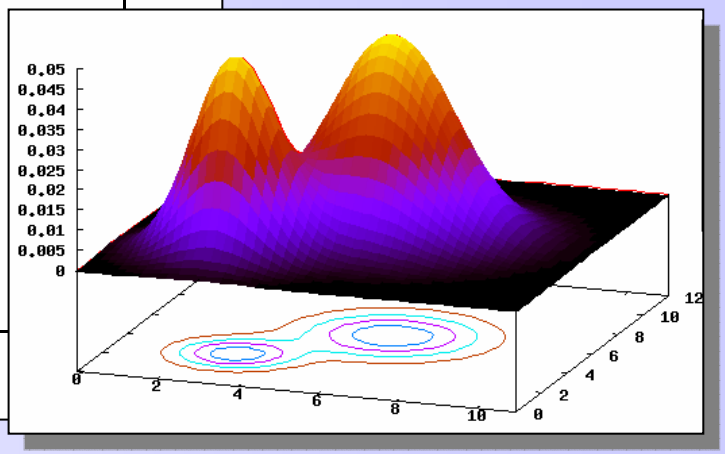
$$x_1^o = y_1^o = 3$$

$$\sigma_{x_1} = \sigma_{y_1} = 1$$

$$x_2^o = y_2^o = 6$$

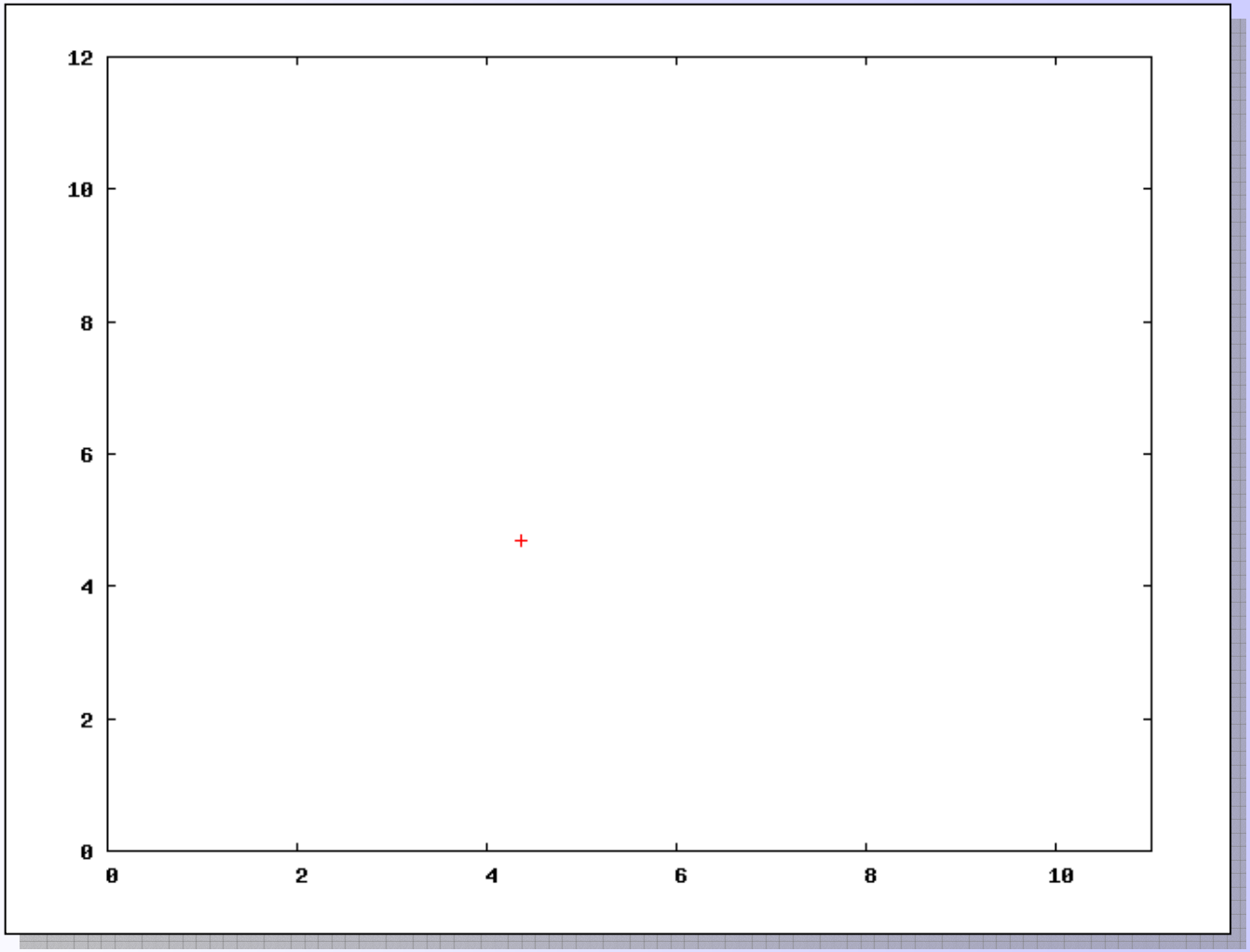
$$\sigma_{x_2} = \sigma_{y_2} = 1.5$$

$$\lambda = 0.7$$

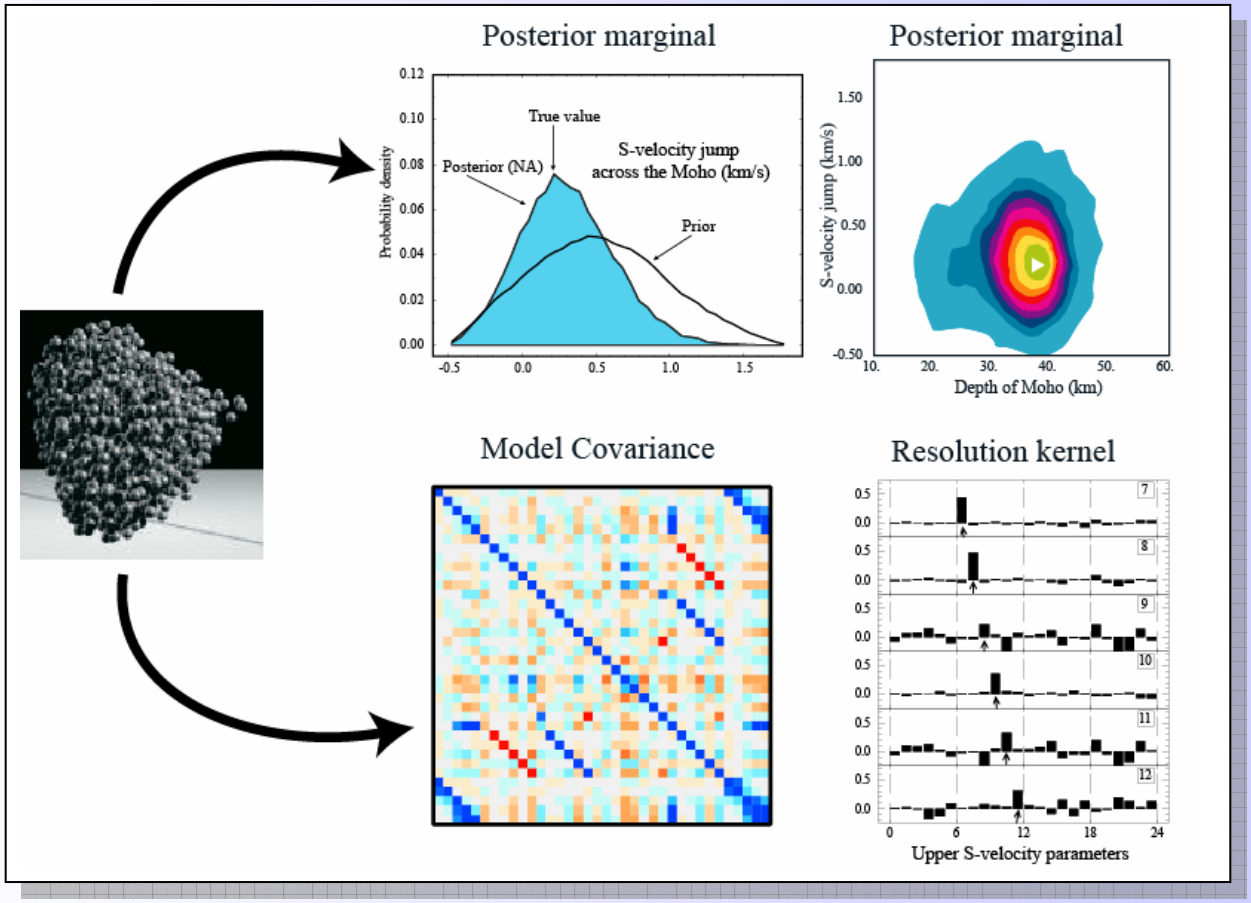


Example: MCMC on a multi-modal PDF

A sum of two Gaussian distributions



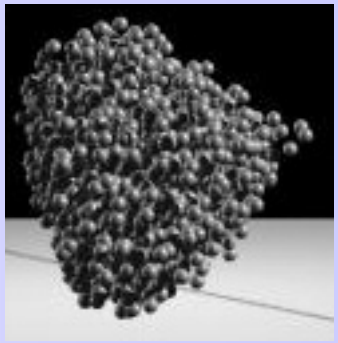
What can we do with samples from the Posterior PDF ?



We are usually interested in things like the model covariance matrix, resolution kernels and marginal probability distributions, and each of these can be represented by an integral (sum over the posterior samples)

$$I = \int_{\mathcal{M}} f(\boldsymbol{m}) p(\boldsymbol{m}|d) d\boldsymbol{m}$$

Monte Carlo integration



Consider any integral of the form

$$I = \int_{\mathcal{M}} f(\boldsymbol{m}) p(\boldsymbol{m}|d) d\boldsymbol{m}$$

Given a set of samples m_i (i=,...,N_s) with sampling density $h(m_i)$, the Monte Carlo approximation to I is given by

$$I \approx \sum_{i=1}^{N_s} \frac{f(\boldsymbol{m}_i) p(\boldsymbol{m}_i|d)}{h(\boldsymbol{m}_i)}$$

Only need to know p(m | d) to a multiplicative constant

If the sampling density is proportional to $p(m_i \mid d)$ then,

$$h(\boldsymbol{m}) = N_s \times p(\boldsymbol{m}|\boldsymbol{d})$$

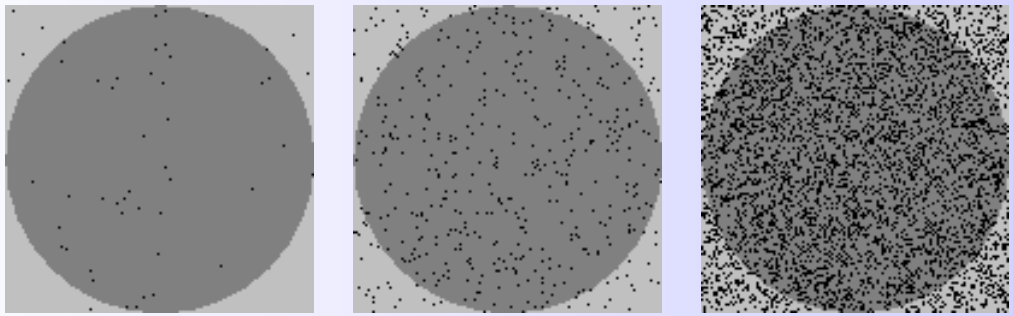
$$\Rightarrow \quad I \approx \frac{1}{N_s} \sum_{i=1}^{N_s} f(\boldsymbol{m}_i)$$

The variance of the $f(m_i)$ values gives the numerical integration error in I

Example: Monte Carlo integration

Finding the area of a circle by throwing darts

$$I = \int_A f(\boldsymbol{m}) d\boldsymbol{m}$$



$$f(\boldsymbol{m}) = \begin{cases} 1 & \boldsymbol{m} \text{ inside circle} \\ 0 & \text{otherwise} \end{cases}$$

$$h(\boldsymbol{m}) = \frac{N_s}{A}$$

$$I \approx \frac{1}{N_s} \sum_{i=1}^{N_S} f(\boldsymbol{m}_i)$$

$$\approx \frac{\text{Number of points inside the circle}}{\text{Total number of points}}$$

Monte Carlo integration

We have

$$I = \int_{\mathcal{M}} f(\boldsymbol{m}) p(\boldsymbol{m}|d) d\boldsymbol{m} \approx \sum_{i=1}^{N_s} \frac{f(\boldsymbol{m}_i) p(\boldsymbol{m}_i|d)}{h(\boldsymbol{m}_i)} \approx \frac{1}{N_s} \sum_{i=1}^{N_s} f(\boldsymbol{m}_i)$$

The variance in this estimate is given by

$$\sigma_I^2 = \frac{1}{N_s} \left\{ \frac{1}{N_s^2} \sum_{i=1}^{N_s} f^2(\boldsymbol{m}_i) - \left(\frac{1}{N_s} \sum_{i=1}^{N_s} f(\boldsymbol{m}_i) \right)^2 \right\}$$

- To carry out MC integration of the posterior we ONLY NEED to be able to evaluate the integrand up to a multiplicative constant.
- As the number of samples, N_s, grows the error in the numerical estimate will decrease with the square root of N_s.
- In principal any sampling density h(m) can be used but the convergence rate will be fastest when h(m) $\propto$ p(m | d).

What useful integrals should one calculate using samples distributed according to the posterior p(m | d)?

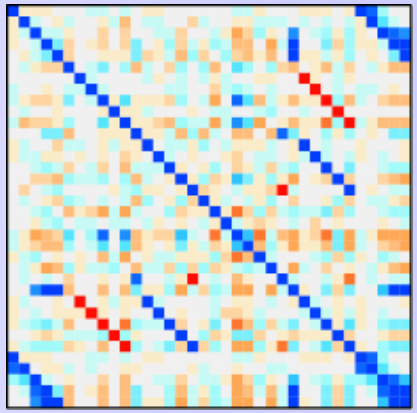
Model covariances from MC integration

The general definition of the Covariance matrix is in terms of a multi-dimensional integral.

First we define the expected value of the i-th parameter

$$E\{m_i\} = \int_{\mathcal{M}} m_i \; p(\boldsymbol{m}|\boldsymbol{d}) d\boldsymbol{m}$$

$$\approx \frac{1}{N_s} \sum_{k=1}^{N_s} m_i^{(k)}$$



$$C_{i,j} = \int_{\mathcal{M}} (m_i - E\{m_i\})(m_j - E\{m_i\}) p(\boldsymbol{m}|\boldsymbol{d}) d\boldsymbol{m}$$

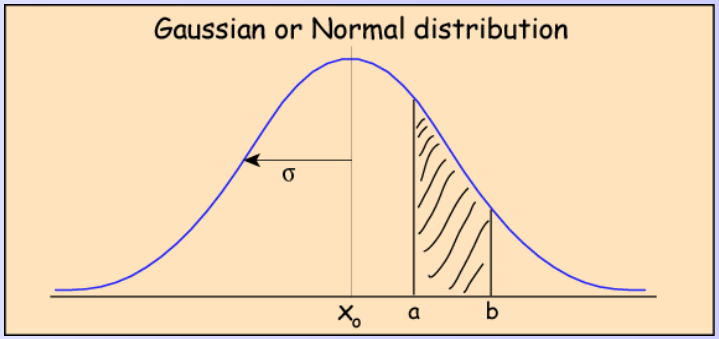
$$C_{i,j} = \int_{\mathcal{M}} m_i m_j p(\boldsymbol{m}|\boldsymbol{d}) d\boldsymbol{m} - E\{m_i\}E\{m_j\}$$

$$\approx \frac{1}{N_s^2} \sum_k m_i^{(k)} m_j^{(k)} - E\{m_i\}E\{m_j\}$$

Example: Model covariances from MC integration

Generate samples whose density is distributed according to a 1-D Gaussian

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x - x_o)^2}{2\sigma^2}\right\}$$



Calculate the expected value for x from the samples (x^1, x^2, ..., X^{Ns})

$$E\{x\} = \frac{1}{N_s} \sum_{k=1}^{N_s} x^{(k)}$$

What value will we get as $N_s \to \infty$?

Calculate the variance of the samples

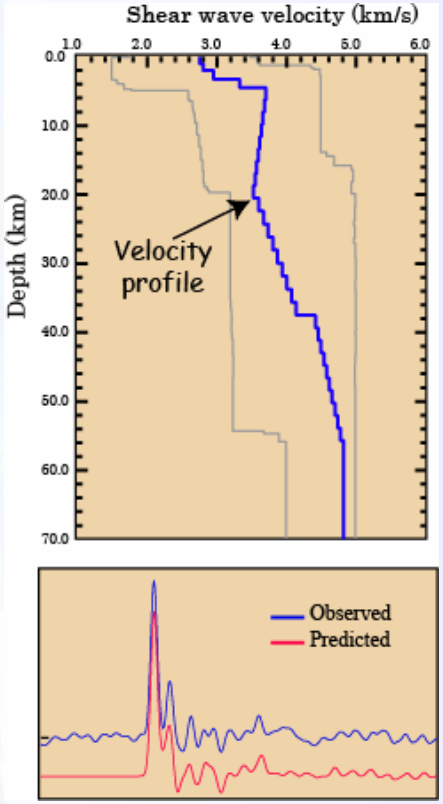
$$C_{1,1} = \frac{1}{N_s} \sum_k (x^{(k)} - E\{x\})^2$$

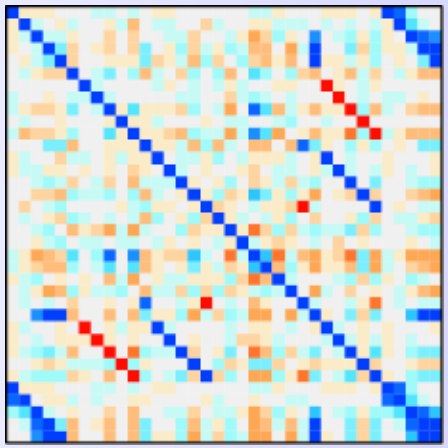
What value will we get as $N_s \to \infty$?

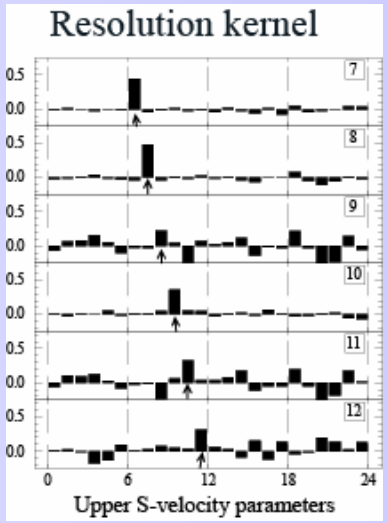
$$C_{1,1} = \frac{1}{N_s^2} \sum_k x^{(k)} x^{(k)} - E\{x\}^2$$

Example: Model covariances from MC integration

Generate samples whose density is distributed according to a 1-D Gaussian







$$R = I - C_{prior}^{-1} C_M$$

MCMC from 10^5 samples

$$C_{1,1} = \frac{1}{N_s^2} \sum_k x^{(k)} x^{(k)} - E\{x\}^2$$

$$E\{x\} = \frac{1}{N_s} \sum_{k=1}^{N_s} x^{(k)}$$

From Sambridge (1999)

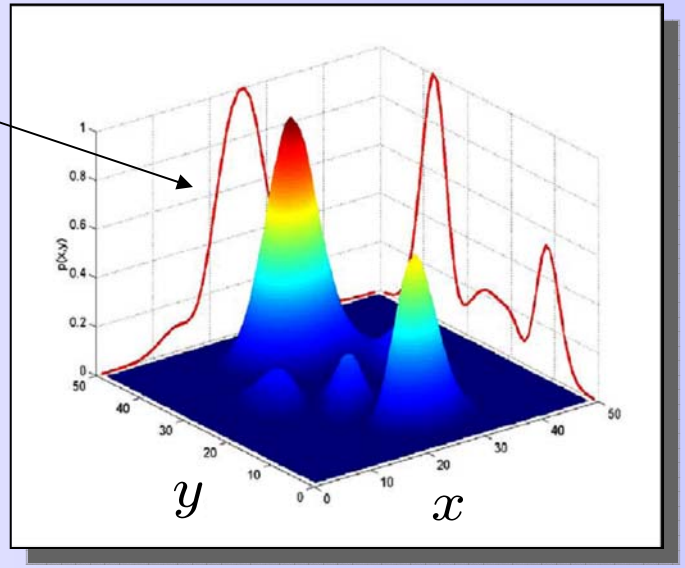
Marginal probability density functions

$$p(x|\boldsymbol{d}) = \int p(x,y|\boldsymbol{d})dy$$

$$p(x,y|\boldsymbol{d}) = \int p(x,y,z|\boldsymbol{d})dz$$

In general

$$p(m_i|\boldsymbol{d}) = \int \ldots \int p(\boldsymbol{m}|\boldsymbol{d}) \prod_{\substack{k=1 \\ k\neq i}}^{N_s} dm_k$$



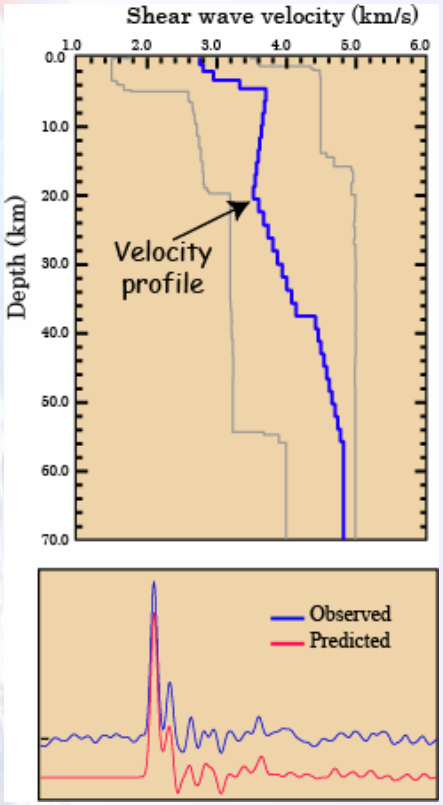
If we have a set of models x_i (i=1,..., N_p) distributed according to the posterior PDF, then we calculate the marginal of x by simply making a histogram of the models as a function of the x co-ordinate only.

Marginals give information about a variable while taking into account likely variations of all other parameters

In the same way from any set of M-dimensional model space vectors we can calculate marginals for a subset of the parameters by simply projecting the vectors onto the lower dimensional space.

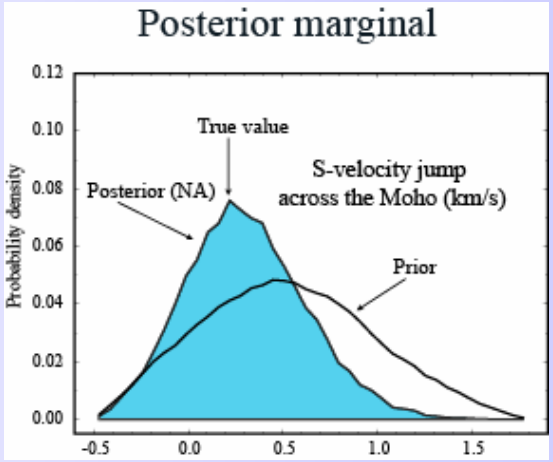
Example: Marginal probability density functions

Receiver function inversion

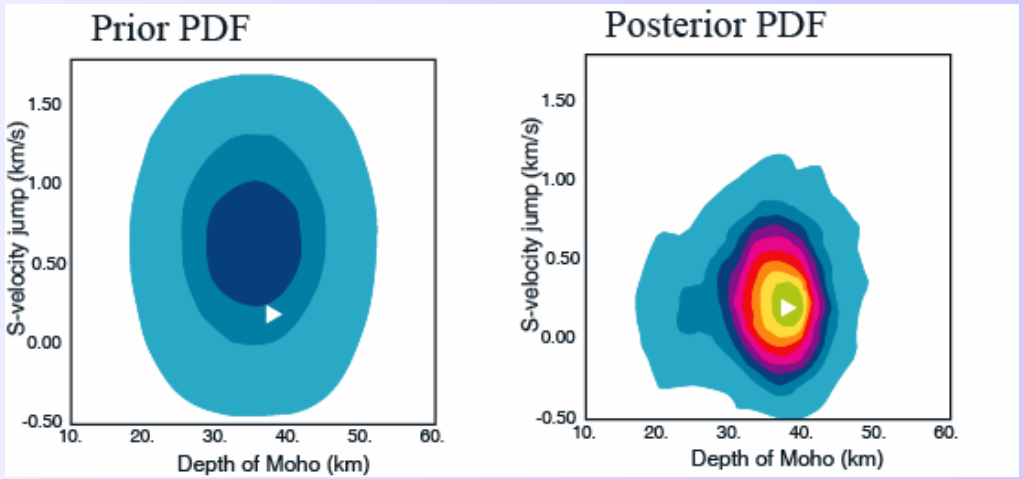


MCMC from 10^5 samples

1-D Marginal



2-D Marginals



From Sambridge (1999)

Example: Marginal probability density functions

Inversion of lunar seismograms from Apollo project 1969-1977

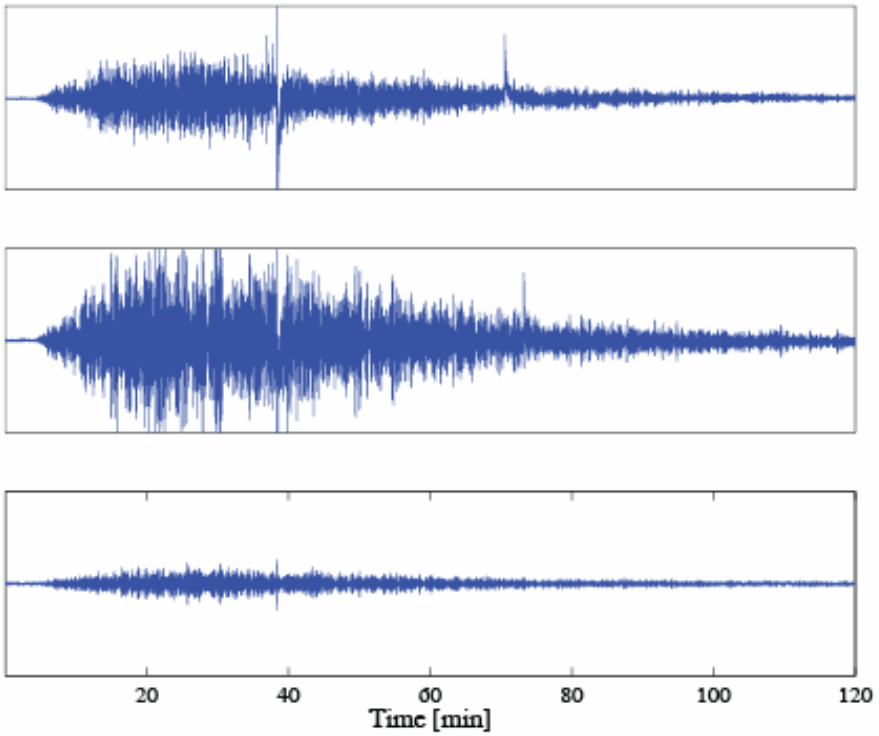


Figure 2. This ?gure shows a Lunar three-component seismic record from a meteoroid impact recorded at station 16 on day 319, 1976. The seismogram at the top is the N–S component (S positive), the middle is E–W (W positive) and the seismogram at the bottom is the vertical component (up positive). The seismograms commence at 23 h 16 min 50 s.

Moonquake from 2/02/1977

Data:
177 First arrival times of moonquakes

Model parameters:
450 unknowns for 1-D P and S-velocity models + hypocentres

From Khan (2000)

Example: Marginal probability density functions

Marginals of P-wave velocity as a function of depth

Prior on velocity parameters

$$\log p(v(z)) = C, \quad a < v(z) < b$$

Likelihood,

$$\propto e^{-\left|d_i^{obs} - g_i(\boldsymbol{m})\right|}$$

MCMC details:

- 1.37 million iterations
- 40-50% acceptance ratio

Marginals were calculated for all velocity parameters and plotted as a function of depth.

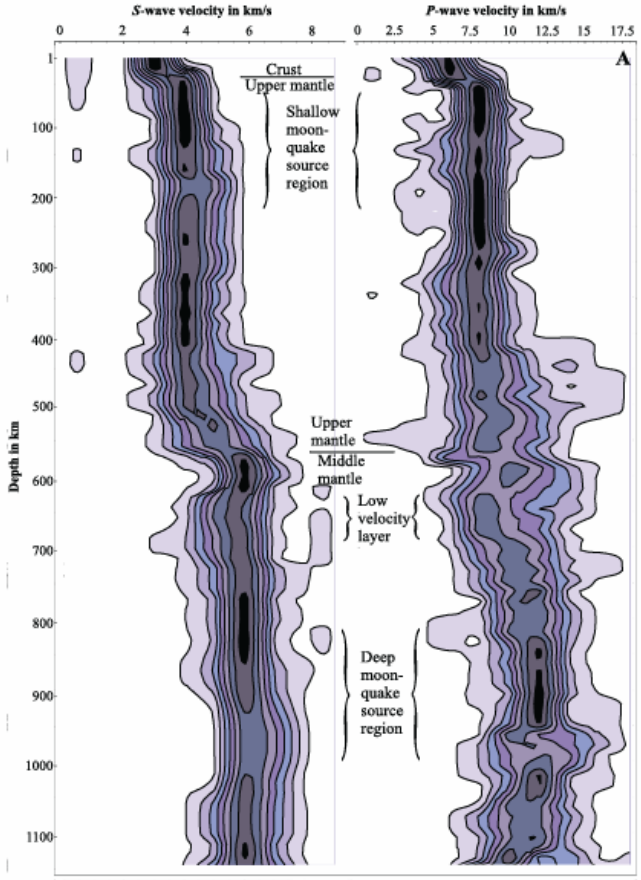


Figure 3. Marginal posterior velocity distributions for the velocity structure of the Moon. 50 000 models were used in constructing the two results. For each kilometre in depth, a histogram re?ecting the marginal *a posteriori* probability distribution of sampled velocities has been computed. These marginals are lined up, and contour lines de?ne nine equal-sized probability density intervals for the distributions.

From Khan (2000)

The End