When one of things that you don‘t know is the number of things that you don‘t know

Malcolm Sambridge

Research School of Earth Sciences,
Australian National University
Canberra, Australia.

Sambridge. M., Gallagher, K., Jackson, A. & Rickwood, P. (2006)
Geophys. J. Int., **167**, 528-542, doi: 10.1111/j.1365246X2006.03155.x.

http://rses.anu.edu.au/~malcolm/papers

Trans-dimensional inverse problems, model comparison and the evidence

Malcolm Sambridge

Research School of Earth Sciences,
Australian National University,
Canberra, Australia.

Sambridge. M., Gallagher, K., Jackson, A. & Rickwood, P. (2006)
Geophys. J. Int., **167**, 528-542, doi: 10.1111/j.1365246X2006.03155.x.

http://rses.anu.edu.au/~malcolm/papers

This talk is about two things

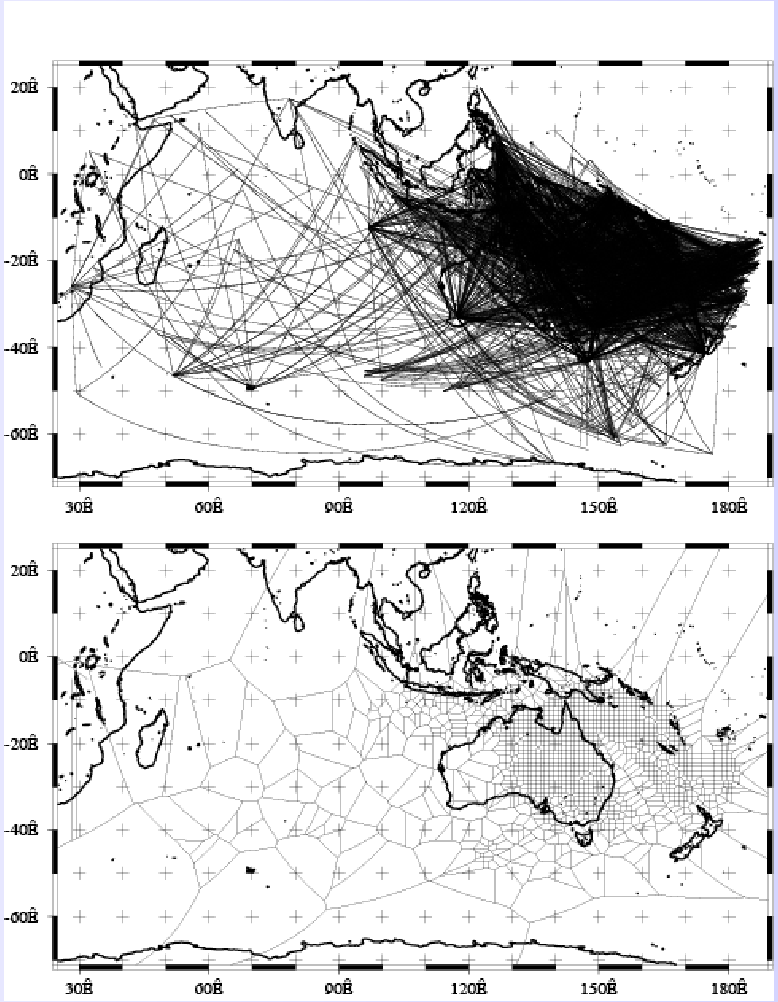
- Trans-dimensional inverse problems

  Those with a variable number of unknowns

- The evidence

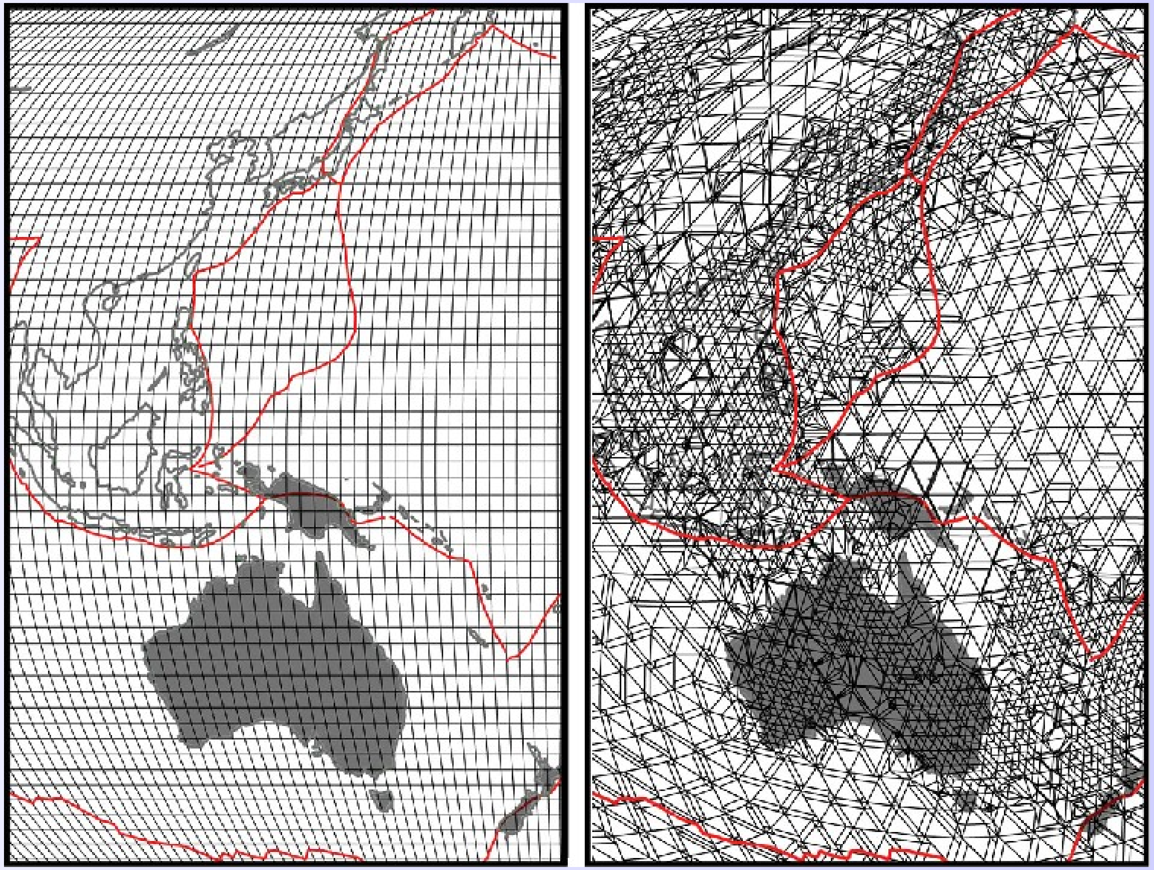
  A quantity from probability theory that allows us to quantitatively compare independent studies performed by different researchers in different institutes at different times

Irregular grids in seismology



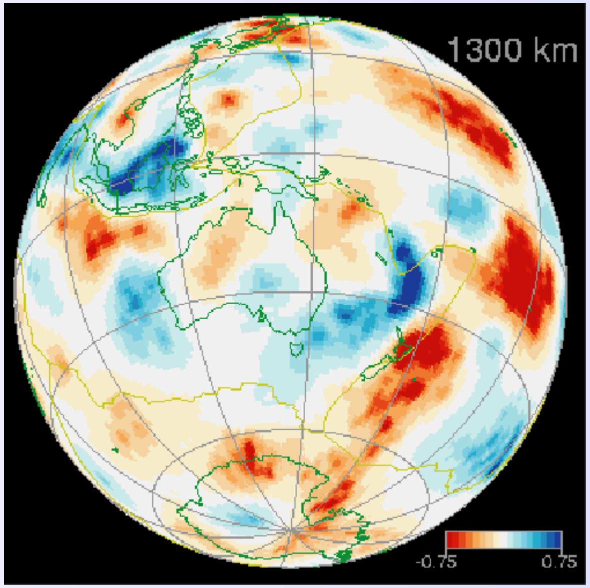
Debayle & Sambridge (2005)

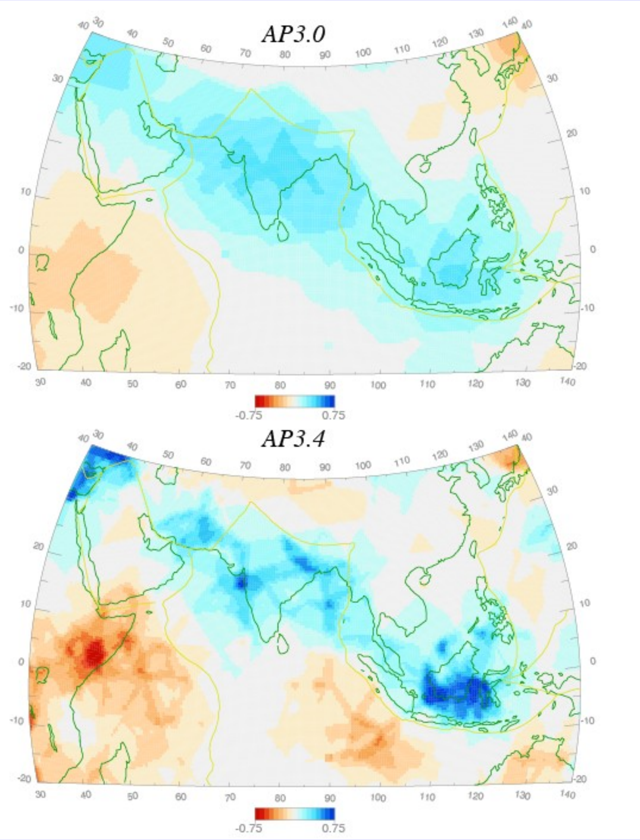
Self Adaptive seismic tomography



Sambridge & Rawlinson (2005)

Self Adaptive seismic tomography

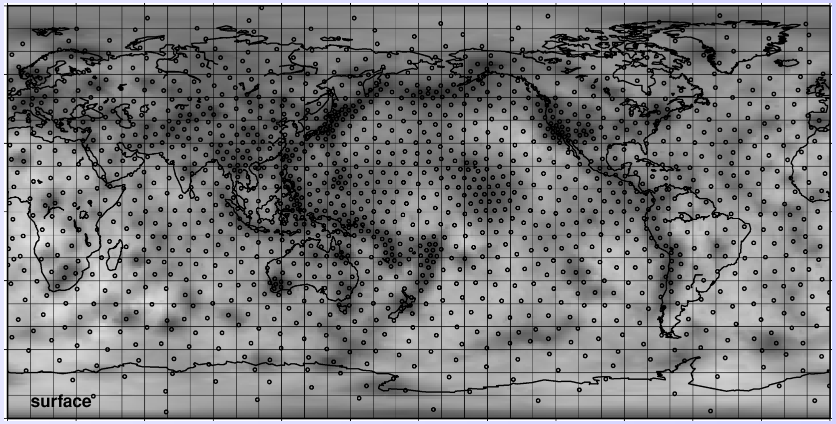


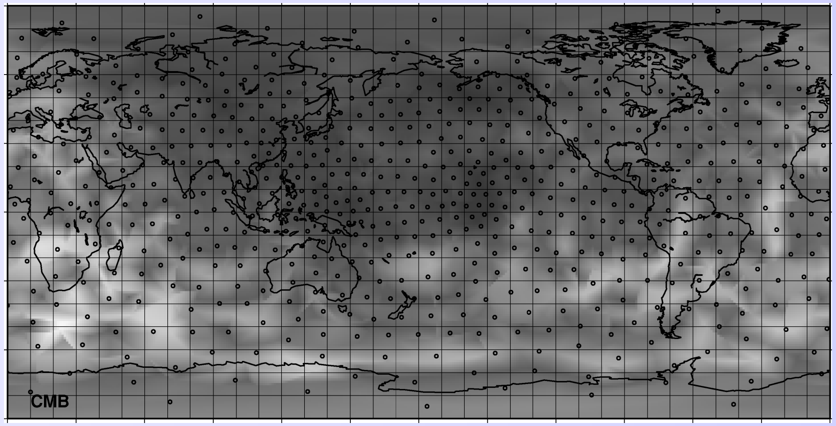


Sambridge & Faletic (2002)

Adaptive grids in seismic tomography

Generating optimized grids from resolution functions





Nolet & Montelli (2005)

Spatially variable grids in tomography

The use of spatially variable parameterizations in seismic tomography is not new….

Some papers on `static' and `dynamic' parameterizations:

Chou & Booker (1979); Tarantola & Nercessian (1984);
Abers & Rocker (1991); Fukao et al. (1992); Zelt & Smith (1992);
Michelini (1995); Vesnaver (1996); Curtis & Snieder (1997);
Widiyantoro & can der Hilst (1998); Bijwaard et al. (1998);
Böhm et al . (2000); Sambridge & Faletic (2003).

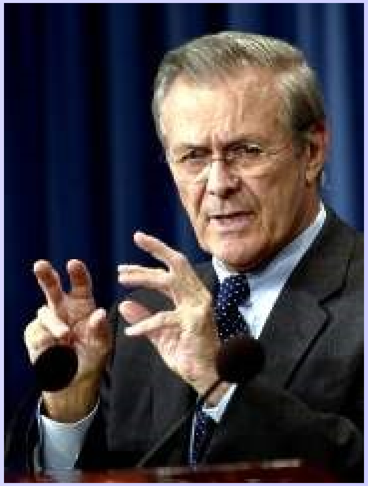
For a recent review see:

``Seismic Tomography with Irregular Meshes'', Sambridge & Rawlinson (Seismic Earth, AGU monograph Levander & Nolet, 2005.)

What is a trans-dimensional inverse problem ?

"As we know, there are known knowns. There are things we know we know. We also know there are known unknowns. That is to say we know there are some things we do not know. But there are also unknown unknowns, the ones we don't know we don't know."

Department of Defense news briefing, Feb. 12, 2002.

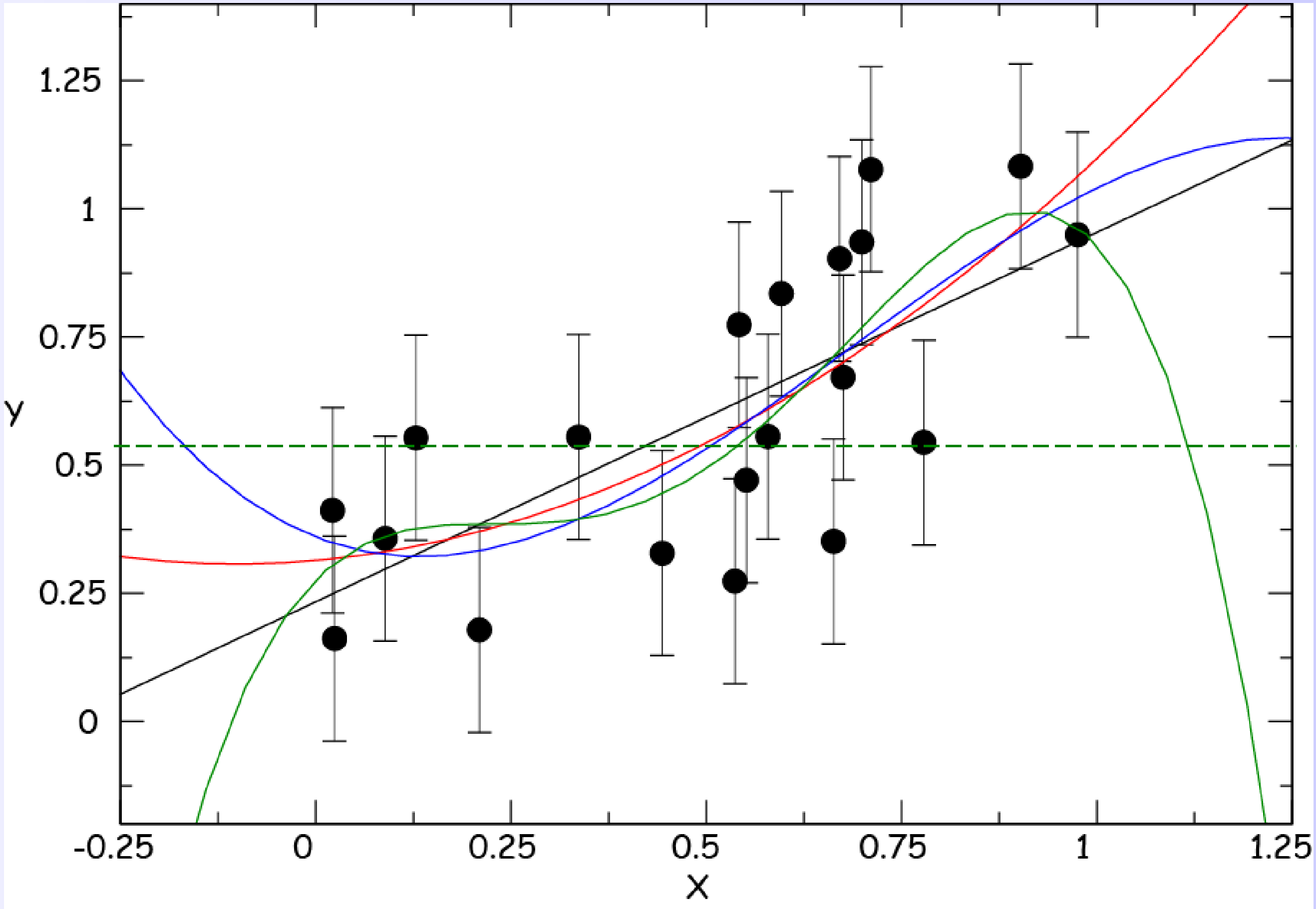


Donald Rumsfeld.

When one of the things you don't know is the number of things you don't know

How many variables required ?

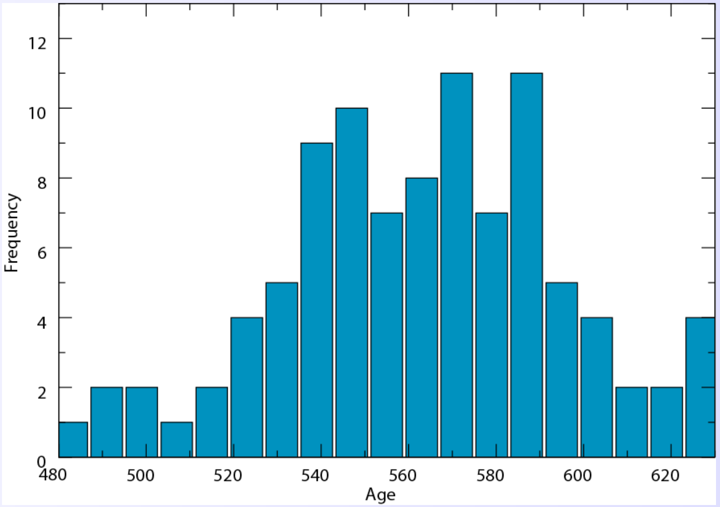
Which curve produced that data ?

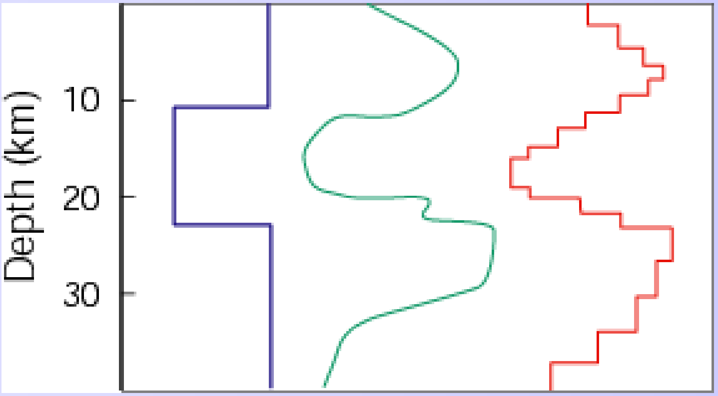


How many parameters should I use to fit my data ?

How many components ?

How many layers ?





This is a *Trans-dimensional* data fitting problem

What is a trans-dimensional inverse problem ?

$$m(x) = \sum_{i=1}^{k} \alpha_i B_i(x)$$

$m(x) =$ Earth model (that we want to recover)

$B_i(x) =$ Basis functions (that are chosen)

$\alpha_i =$ Coefficients (unknowns)

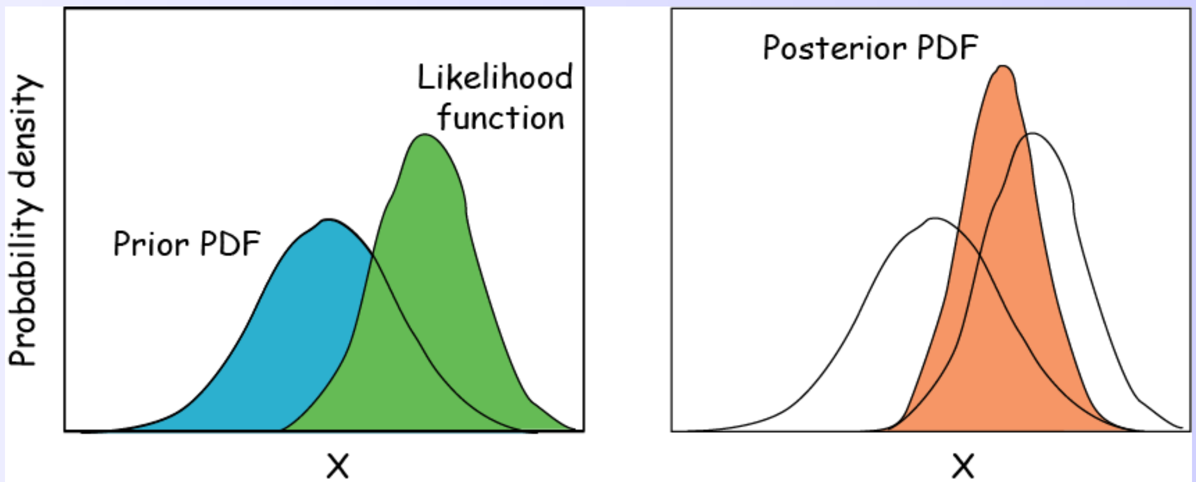
$k =$ The number of unknowns (unknown)

This is a hierarchical parametrization

1702-1761

Probabilistic inference

All information is expressed in terms of probability density functions



Bayes' rule

Conditional PDF

All additional assumptions are here

$$p(m\,|\,d,H) \propto p(d\,|\,m,H) \times p(m,H)$$

A posteriori probability density $\propto$ Likelihood x a priori probability density

The evidence

The evidence $p(d \mid H)$ is also known as the marginal likelihood

$$p(d \mid H) = \int p(d \mid m, H)\, p(m \mid H)\, dm$$

$$p(m \mid d, H) = \frac{p(d \mid m, H)\, p(m \mid H)}{p(d \mid H)}$$

$$Posterior = \frac{Likelihood \times prior}{Evidence}$$

Geophysicists have often overlooked it because it is not a function of the model parameters. It measures the fit of the theory !

J. Skilling, however, describes it as the `*single most important quantity in all of Bayesian inference*'. This is because it is transferable quantity between independent studies.

If we all published evidence values of our model fit then studies could be quantitatively compared !

Model choice and Occam's razor

`A theory with mathematical beauty is more likely to be correct than an ugly one that fits the some experimental data' – Paul Dirac

Occam's razor suggests we should prefer the simpler theory which explains the data

Suppose we have two different theories H_1 and H_2

Plausibility ratio

$$\frac{p(H_1 \mid d)}{p(H_2 \mid d)} = \frac{p(d \mid H_1)}{p(d \mid H_2)} \frac{p(H_1)}{p(H_2)}$$

Model predictions (Bayes factor) or ratio of evidences

Prior preferences

This tells us how well the data support each theory

Model choice: An example

What are the next two numbers in the sequence ?

-1, 3, 7, 11, ?, ?

$H_1:$ 15, 19 because $x_{i+1} = x_i + 4$

But what about the alternate theory ?

$H_2:$ -19.9, 1043.8 because $x_{i+1} = -x_i^3/11 + 9x_i^2/11 + 23/11$

$$\frac{p(H_1 \mid d)}{p(H_2 \mid d)} = \frac{p(d \mid H_1)}{p(d \mid H_2)} \frac{p(H_1)}{p(H_2)}$$

Let us assume we have no prior preference for either theory $p(H_1) = p(H_2)$

Model comparison: An example

$H_1:$ An arithmetic progression fits the data $x_{i+1} = x_i + a$ (2 parameters)

$H_2:$ A cubic sequence fits the data $x_{i+1} = cx_i^3 + dx_i^2 + e$ (4 parameters)

We must find the evidence ratio $$\frac{p(d \mid H_1)}{p(d \mid H_2)}$$

$p(d \mid H_1)$ Is called the evidence for model 1 and is obtained by specifying the probability distribution each model assigns to its parameters

If a and x_o equally likely in range [-50,50] then

$$p(d \mid H_1) = \frac{1}{101}\frac{1}{101} = 0.0001$$

If c, d, and e have numerators in [-50,50] and denominators in [1,50]

$$p(d \mid H_2) = \frac{1}{101}\frac{4}{101}\frac{1}{50}\frac{4}{101}\frac{1}{50}\frac{2}{101}\frac{1}{50} = 2.5 \times 10^{-12}$$

40 million to 1 in favour of the simpler theory !

See Mackay (2003)

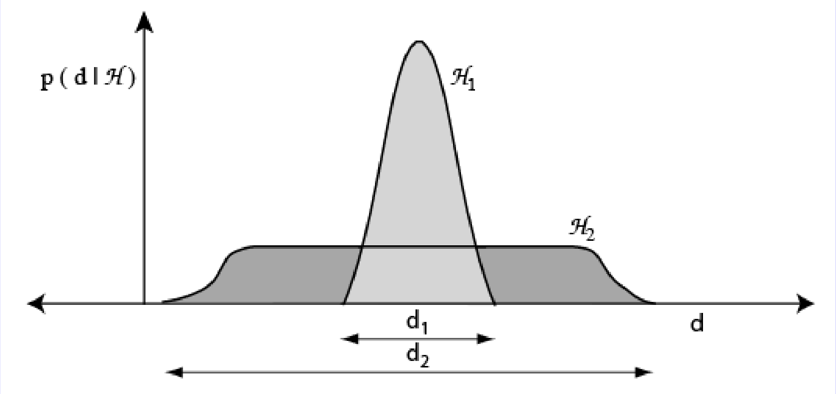
The natural parsimony of Bayesian Inference

Without a prior preference expressed for the simpler model, Bayesian inference *automatically* favours the simpler theory with the fewer unknowns (provided it fits the data).

$$p(d \mid H) = \int p(d \mid m, H)\, p(m \mid H)\, dm$$

Evidence measures how well the theory fits the data

Bayesian Inference rewards models in proportion to how well they predict the data. Complex models are able to produce a broad range of predictions, while simple models have a narrower range.

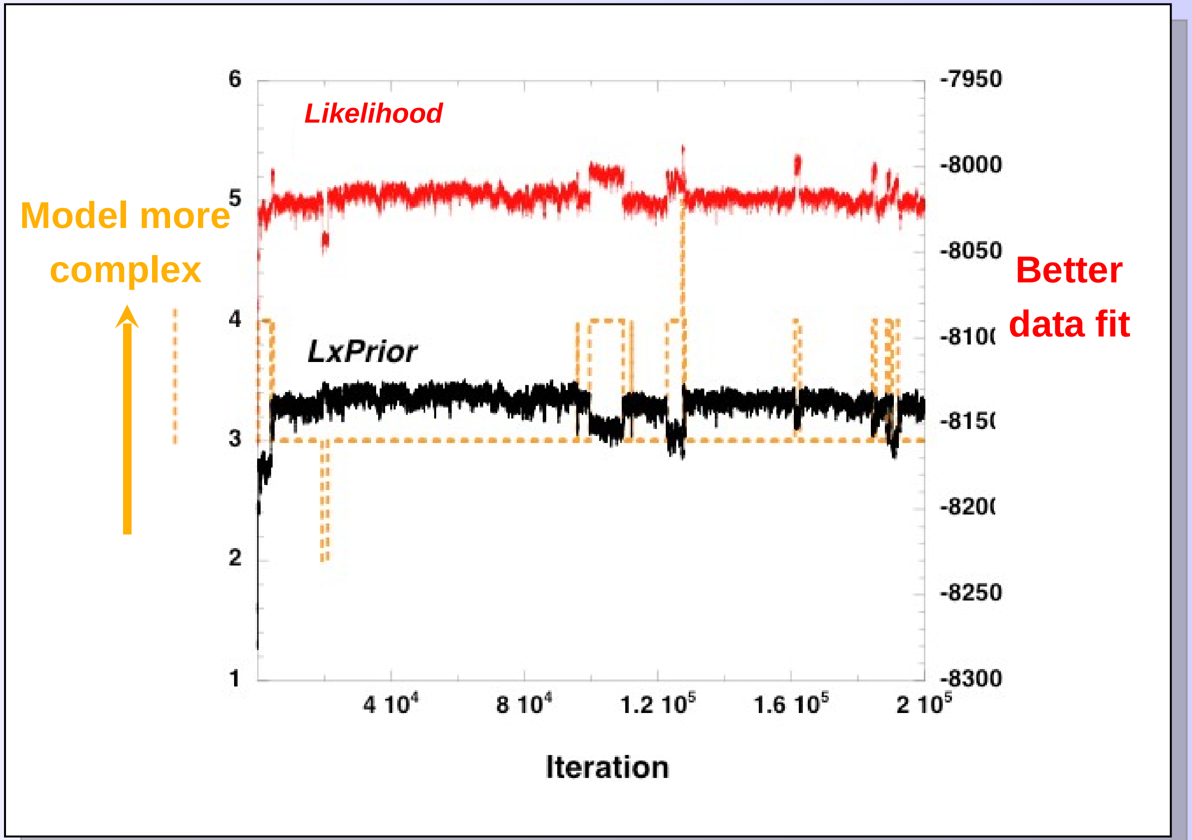


If both a simple and complex model fit the data then the simple model will predict the data more strongly in the overlap region

From Mackay (2003)

Natural parsimony of Bayesian Inference

Bayesian’s prefer simpler models



From Thermochronology study of Stephenson, Gallagher & Holmes (EPSL, 2006)

Trans-dimensional inverse problems

Consider an inverse problem with *k* unknowns

Bayes' rule for the model parameters

$$p(m \mid d, k) = \frac{p(d \mid m, k)\, p(m \mid k)}{p(d \mid k)}$$

Bayes' rule for the number of parameters

$$p(k \mid d) = \frac{p(d \mid k)\, p(k)}{p(d)}$$

By combining we get …

Bayes' rule for both the parameters and the number of parameters

$$p(m, k \mid d) = \frac{p(d \mid m, k)\, p(m \mid k)\, p(k)}{p(d)}$$

Can we sample from trans-dimensional posteriors ?

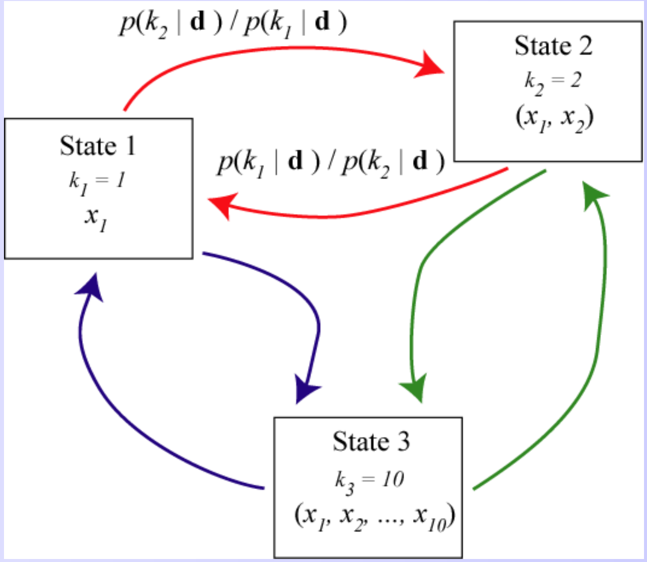
The reversible jump algorithm

A breakthrough was the reversible jump algorithm of Green(1995), which can simulate from arbitrary trans-dimensional PDFs.

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

This is effectively an extension to the well known Metropolis algorithm with acceptance probability

$$\alpha = \mathrm{Min}\left\{1, \frac{p(d\,|\,m_2,k_2)\,p(m_2\,|\,k_2)\,p(k_2)q(m_1\,|\,m_2)}{p(d\,|\,m_1,k_1)\,p(m_1\,|\,k_1)\,p(k_1)q(m_2\,|\,m_1)}\,|\,J\,|\right\}$$



Detailed balance

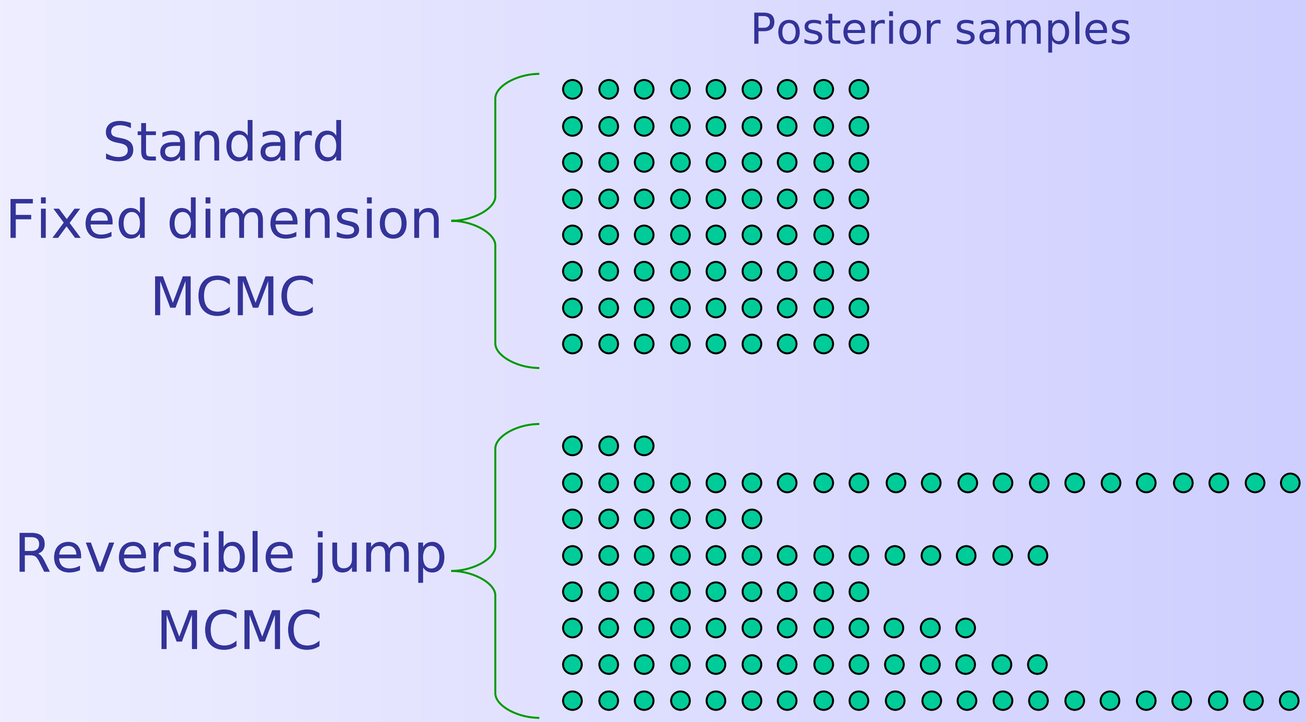
Jacobian $|\,J\,|$ is often, but not always, 1. *Automatic* (Jacobian) implementation of Green (2003) is convenient.

Research focus is on efficient (automatic) implementations for higher dimensions.

See Denision et al. (2002); Malinverno (2002), Sambridge et al. (2006)

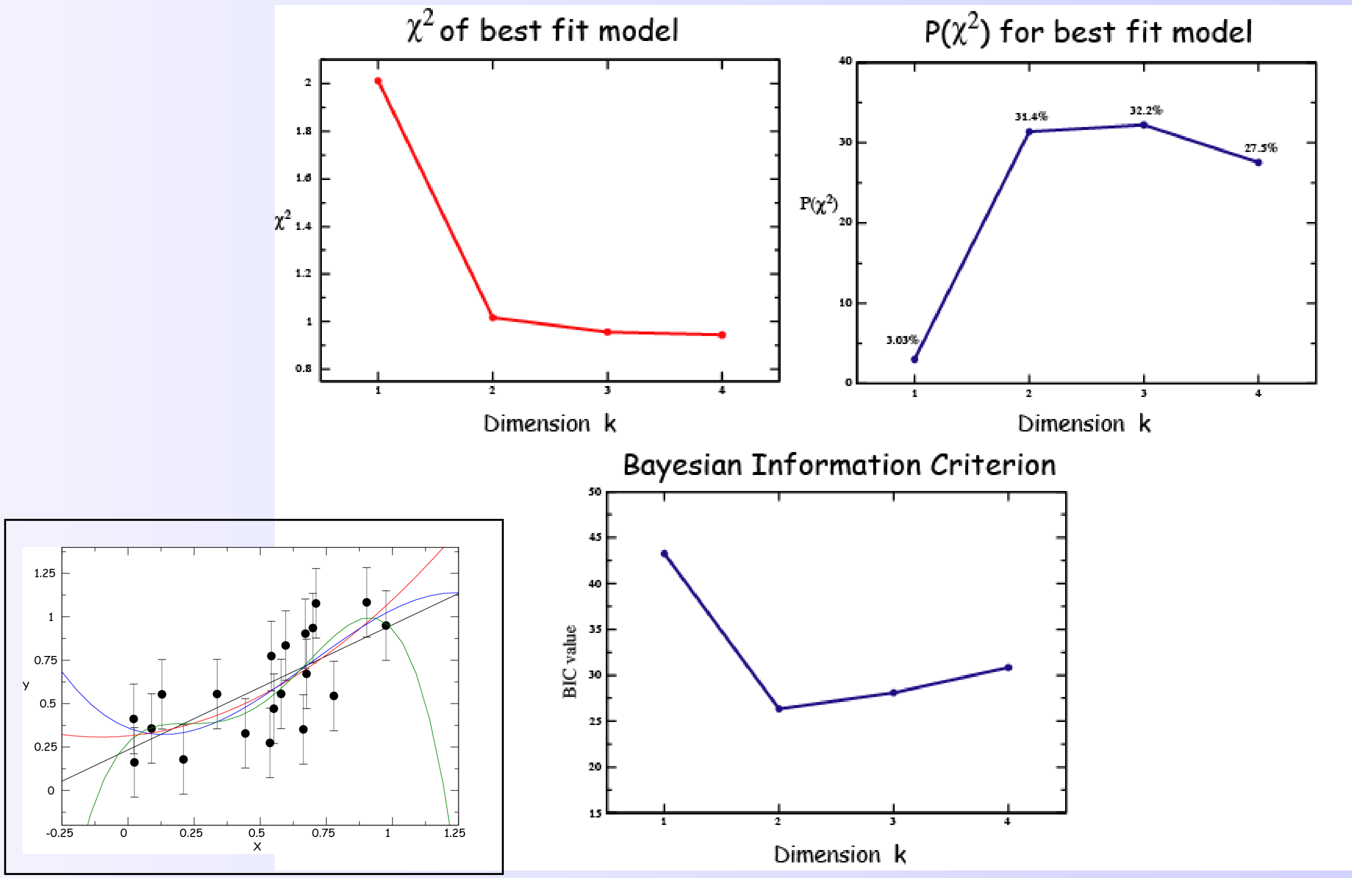
Trans-dimensional sampling

The reversible jump algorithm produces samples from the variable dimension posterior PDF. Hence the samples are models with different numbers of parameters



Trans-dimensional sampling: A regression example

Standard point estimates give differing answers



Reversible jump applied to the regression example

Use the reversible jump algorithm to sample from the trans-dimensional prior and posterior

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

For the regression problem k=1,…,4, and RJ-MCMC produces

State 1: $y = a_{1,1}$ $\quad (a_{1,1})_j \quad (j = 1,\ldots,N_1)$

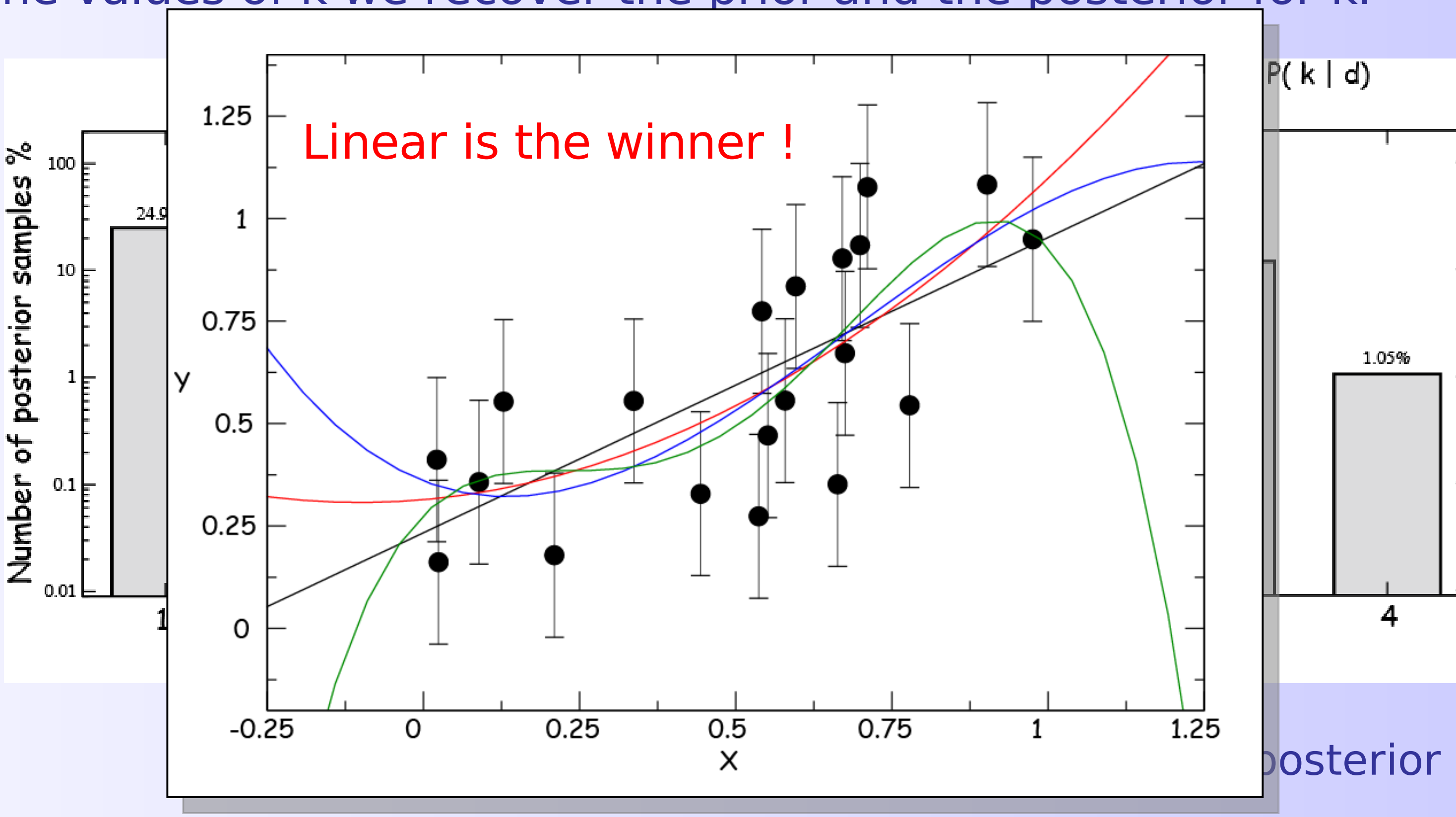
State 2: $y = a_{1,2} + a_{2,2}x$ $\quad (a_{1,2}, a_{2,2})_j \quad (j = 1,\ldots,N_2)$

State 3: $y = a_{1,3} + a_{2,3}x + a_{3,3}x^2$ $\quad (a_{1,3}, a_{2,3}, a_{3,3})_j \quad (j = 1,\ldots,N_3)$

State 4: $y = a_{1,4} + a_{2,4}x + a_{3,4}x^2 + a_{4,4}x^3$ $\quad (a_{1,4}, a_{2,4}, a_{3,4}, a_{4,4})_j \quad (j = 1,\ldots,N_4)$

Reversible jump results: regression example

The reversible jump algorithm can be used to generate samples from the trans-dimensional prior and posterior PDFs. By tabulating the values of k we recover the prior and the posterior for k.



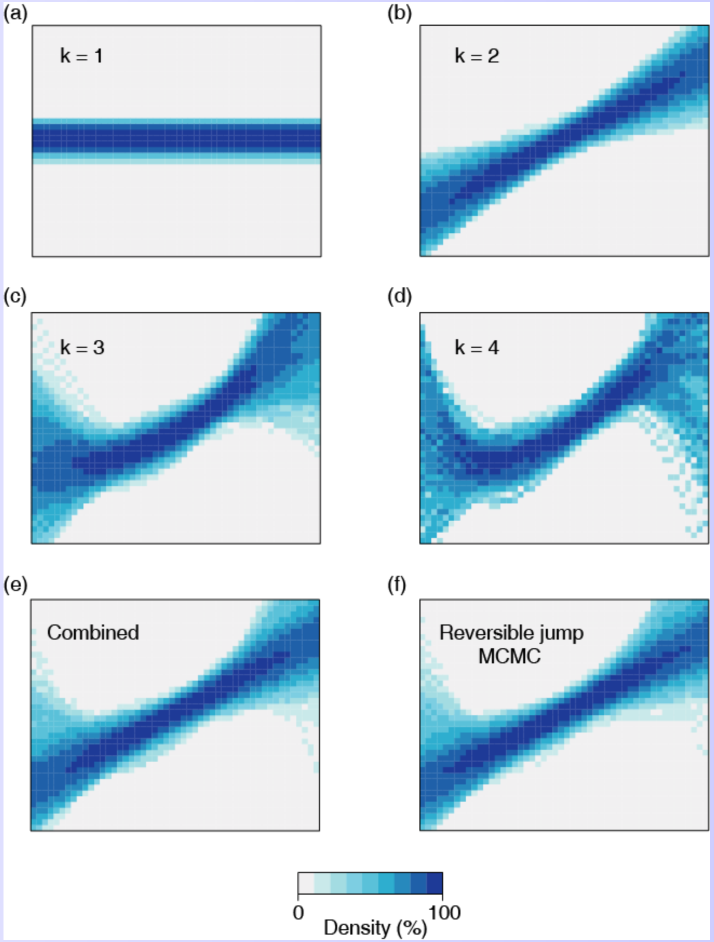
Trans-dimensional sampling from a fixed dimensional sampler

Can the reversible jump algorithm be replicated with more familiar fixed dimensional MCMC ?

Answer: Yes !

1. First generate fixed dimensional posteriors
2. Then combine them with weights P_k

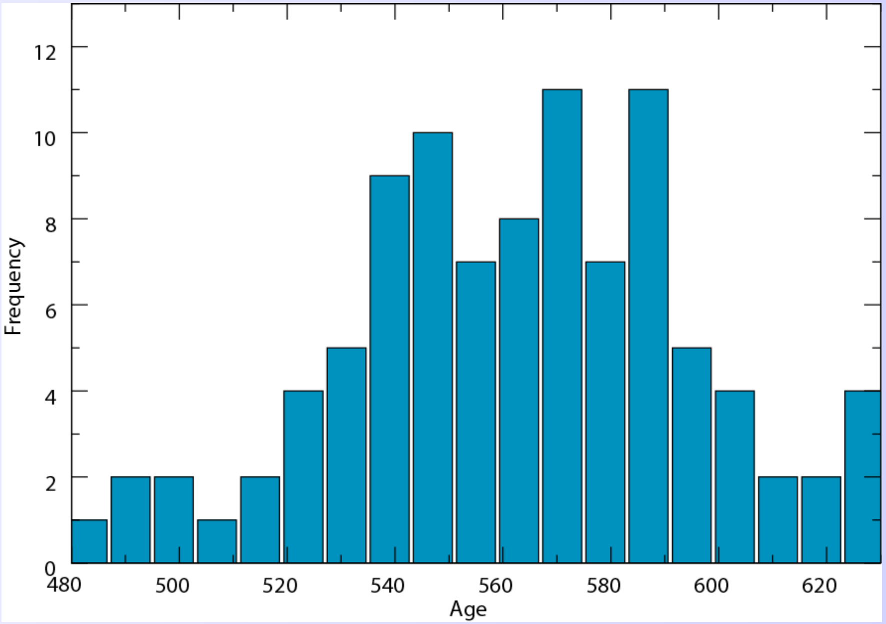
$$P_k = \frac{p(d\,|\,k)\,p(k)}{\sum_{k'}^{k\max} p(d\,|\,k')\,p(k')}$$



Only requires relative evidence values !

From Sambridge et al. (2006)

Trans-dimensional sampling: A mixture modelling example



How many Gaussian components ?

From Sambridge et al. (2006)

Trans-dimensional sampling: A mixture modelling example

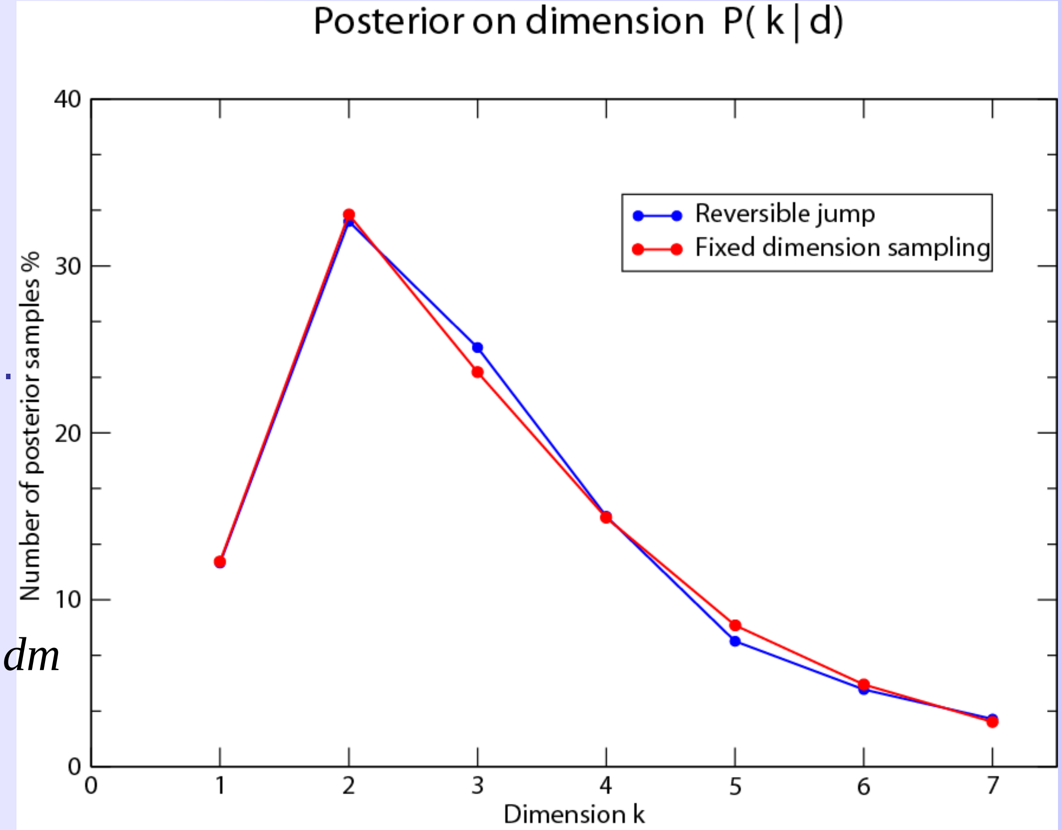
Posterior simulation

- Using reversible jump algorithm,
- Fixed k sampling with evidence weights.

$$p(k \mid d) \propto p(d \mid k)\, p(k)$$

$$p(d \mid k) = \int p(d \mid m, k)\, p(m \mid k)\, dm$$

$$p(d \mid k) \approx \frac{1}{N_k} \sum_{i=1}^{N_k} p(d \mid m, k)$$



Trans-dimensional inversion
Recent applications

- Climate histories

 Inferring ground surface temperature histories from
 high precision borehole temperatures. *(Hopcroft et al. 2008)*

- Thermochronology

 Inferring cooling or erosion histories from spatially distributed Fission
 track data. *(Gallagher et al., 2005, 2006, Stephenson et al. 2006)*

- Stratigraphic modelling

 Using borehole data on grain size and sediment thickness to infer sea-
 level change and sediment flux in a reservoir -> uncertainty on oil
 production rates. *(Charvin et al. 2008)*

- Geochronology

 Inferring number and ages of distinct geological events from
 mixtures of rock ages. *(Jasra et al., 2006)*

Calculating the evidence

But how easy is the evidence to calculate ?

- For small discrete problems it can be easy as in the previous example
- For continuous problems with k < 50. Numerical integration is possible using samples, x_i generated from the prior $p(m \mid H)$

$$p(d \mid H) \approx \frac{1}{N_s} \sum_{i=1}^{N_s} p(d \mid x_i, H) \approx \frac{1}{\sum_{i=1}^{N_s} [p(d \mid x_i, H)]^{-1}}$$

- For some linear (least squares) problems, analytical expressions can be found

$$p(d \mid H) = \frac{\rho(d, \hat{x}, x_o)}{2\pi^{(N_d - k)/2}} \left(\frac{|C_{post}|}{|C_{data}||C_{prior}|} \right)^{1/2}$$

(Malinverno, 2002)

- For large k and highly nonlinear problems – forget it !

See Sambridge et al. (2006) for details

Conclusions

- Trans-dimensional inverse problems are a natural extension to the fixed dimension Bayesian inference familiar to geoscientists.

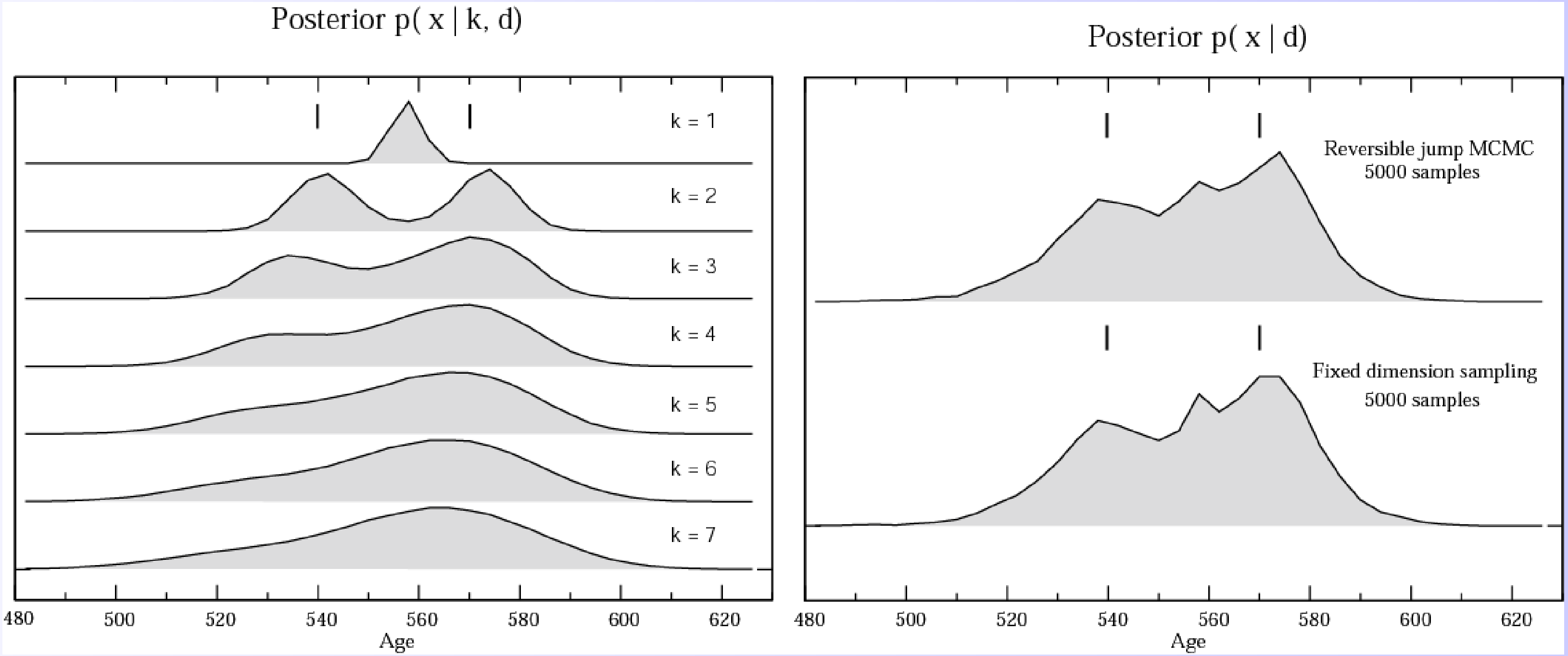
- The ratio of the evidences for fixed dimensions is the key quantity that allows fixed dimensional MCMC tools to be used in variable $p(d \mid k)$ dimensions. (Conversely RJ-MCMC can be used to calculate the evidence

- The evidence is a transferable quantity that allows us to compare apples with oranges.

- Evidence calculations are possible in a range of classes. Transdimensional inverse problems are beginning to find applications in the Earth Sciences.

- Even astronomers are calculating the evidence so why don't we !

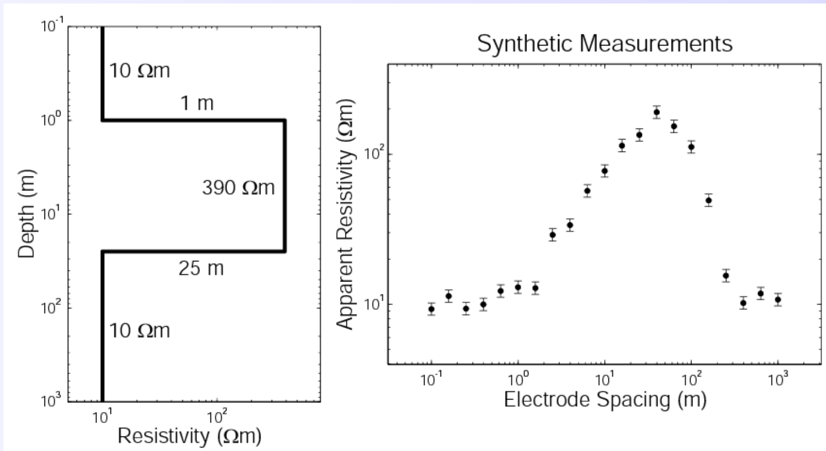
Trans-dimensional sampling:
A mixture modelling example



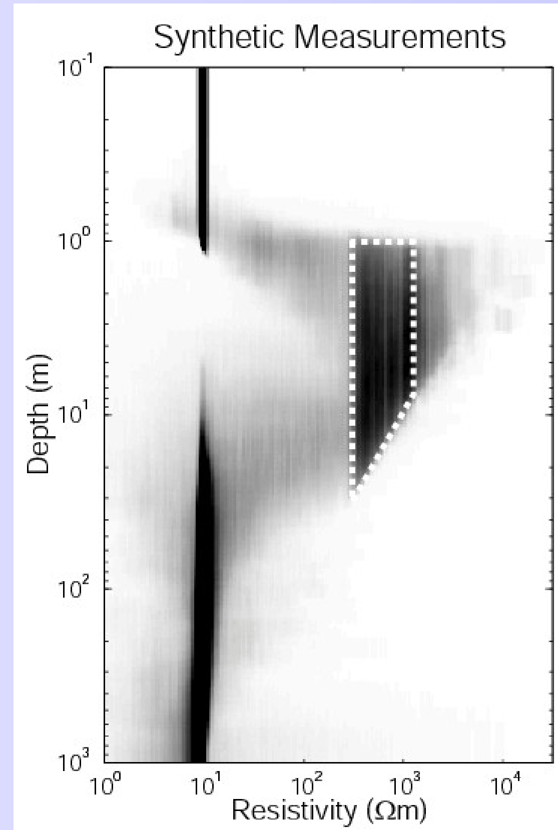
Fixed dimension MCMC simulations

Reversible jump and weighted fixed k sampling

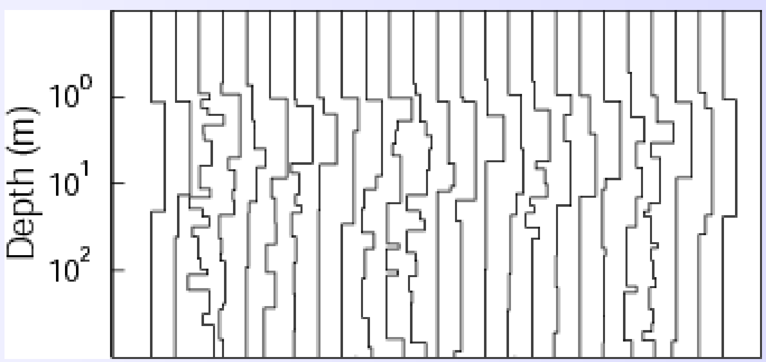
An electrical resistivity example



True model Synthetic data



Posterior density



50 profiles found with posterior sampler

From Malinverno (2002)

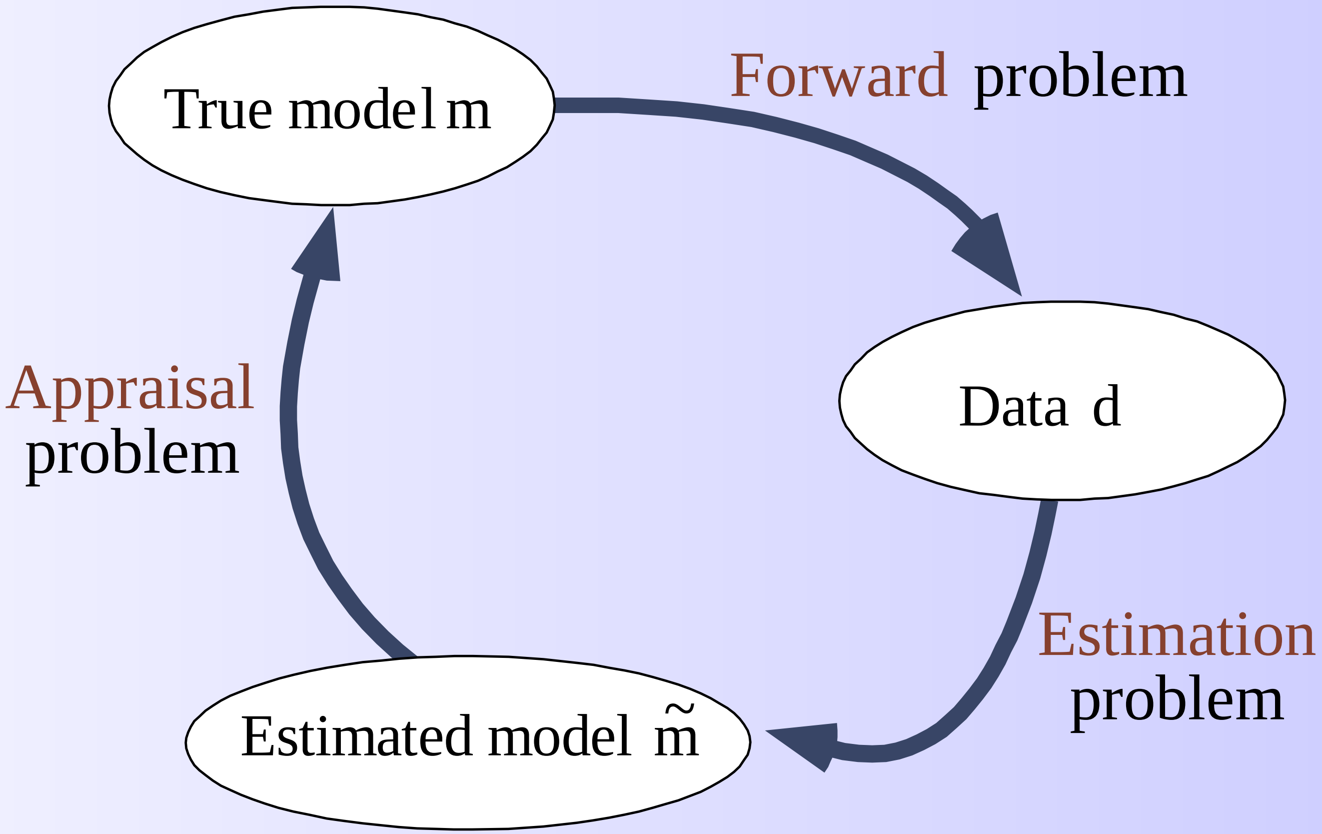
Elements of an Inverse problem

What you get out depends on what you put in:

- The data you have (and its noise)
- The physics of the forward problem
- Your choice of parameterization
- Your definition of a solution

A way of asking questions of data

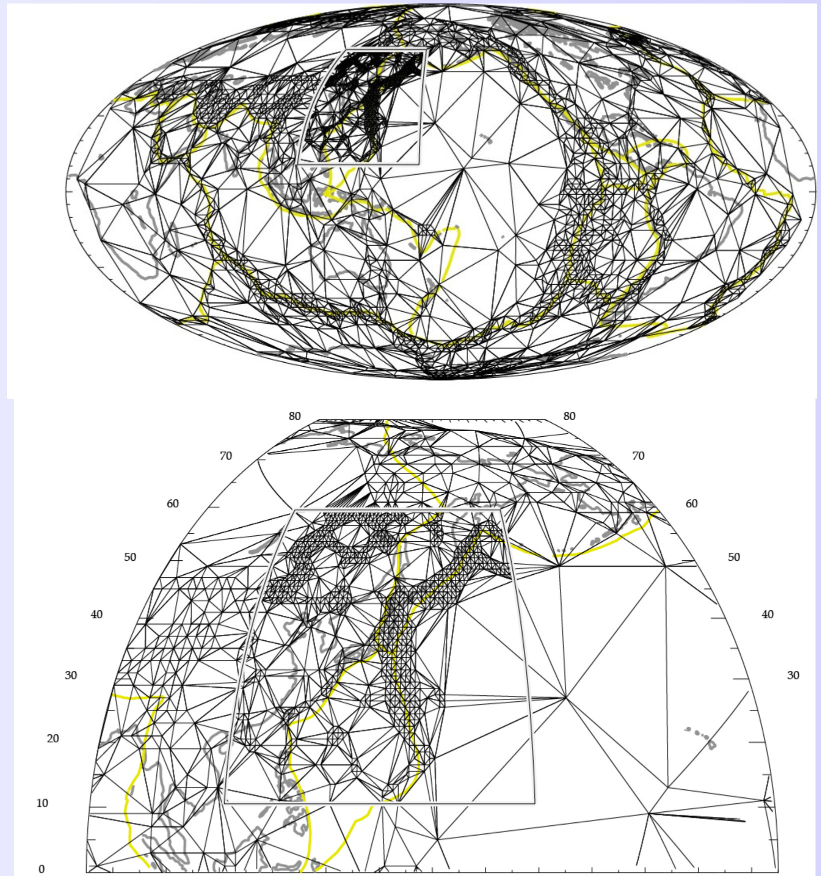
What is an inverse problem ?



From Snieder & Trampert (2000)

Irregular grids in seismology

Case example



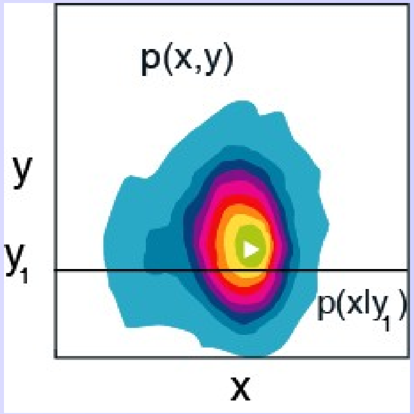
From Sambridge, Braun & McQueen (1995)

Probability density functions

Mathematical preliminaries:

Joint and conditional PDFs

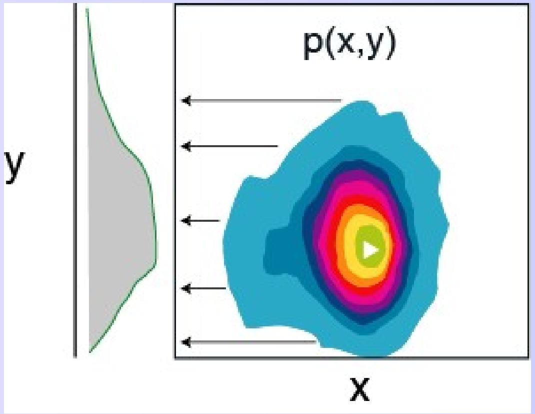
$$p(x, y) = p(x \mid y) \times p(y)$$



Marginal PDFs

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

$$p(x, y) = \int p(x, y, z) dz$$



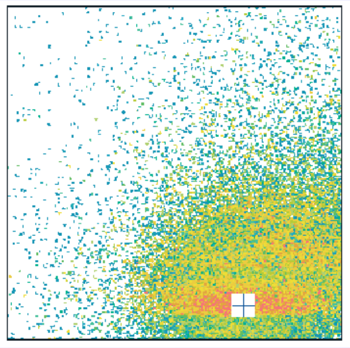
What do we get from Bayesian Inference ?

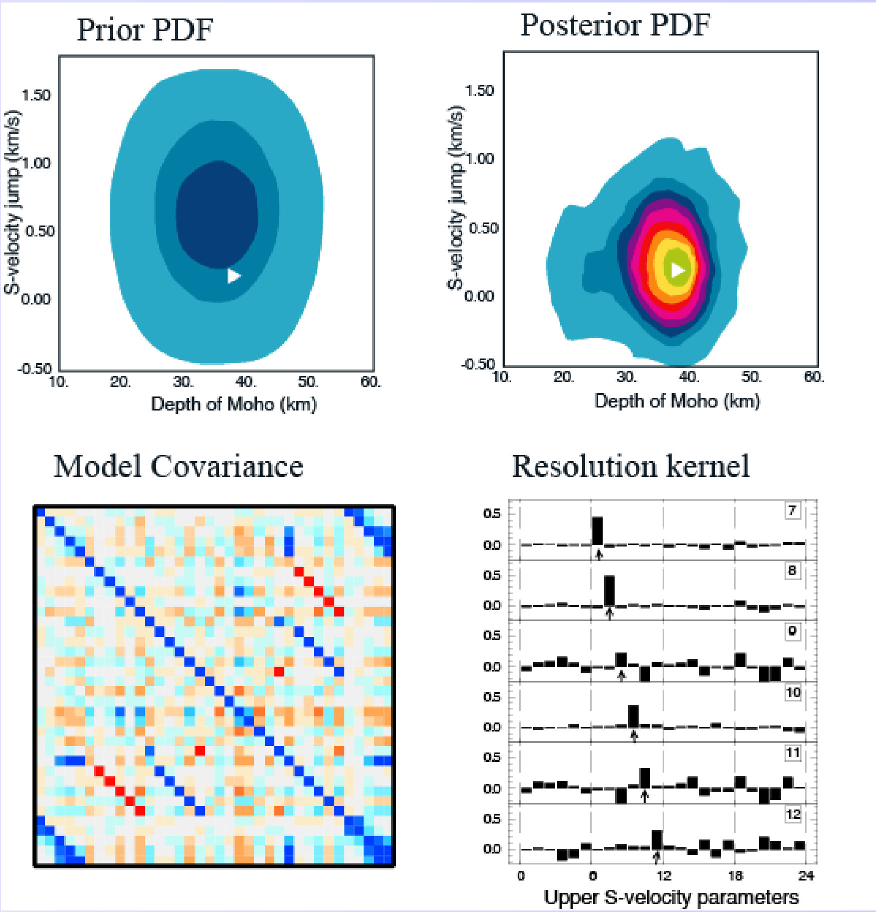
Generate samples whose density follows the posterior

$$\rho(m) \propto p(d \mid m)\, p(m)$$

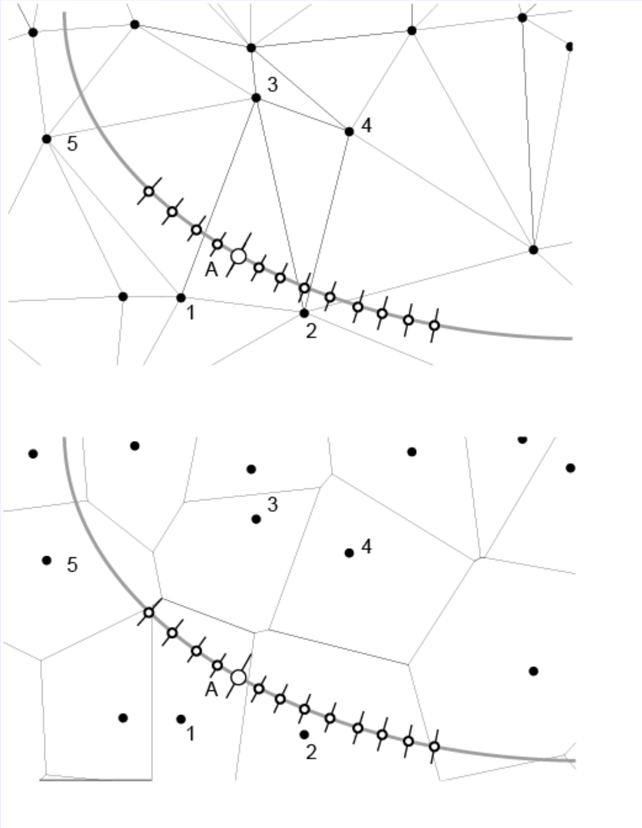
The workhorse technique is Markov chain Monte Carlo

e.g. Metropolis, Gibbs sampling

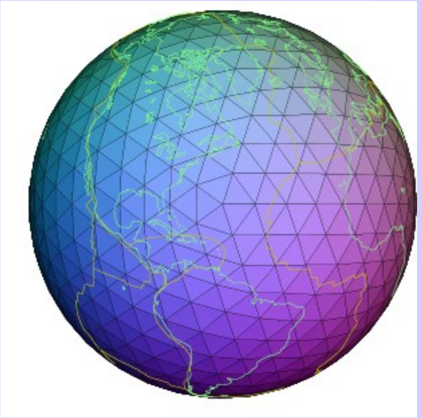


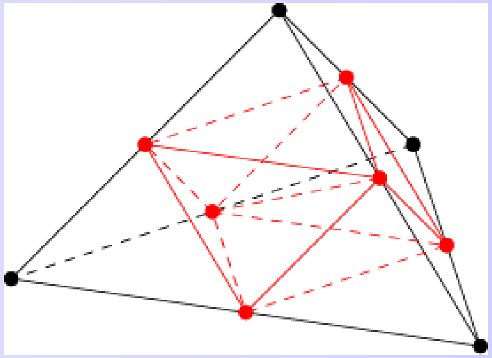


Self Adaptive seismic tomography



From Sambridge & Gudmundsson (1998)

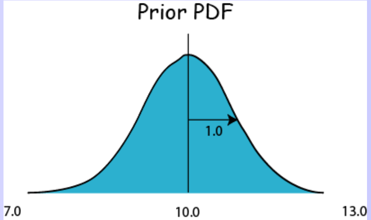




From Sambridge & Faletic (2003)

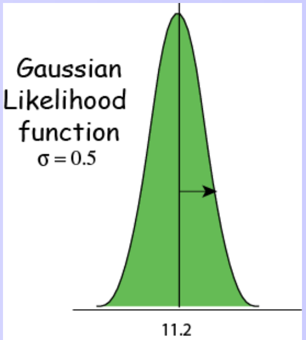
Example: Measuring the mass of an object

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 § 1μ g. In probabilistic terms we could represent this as a Gaussian prior distribution



prior

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



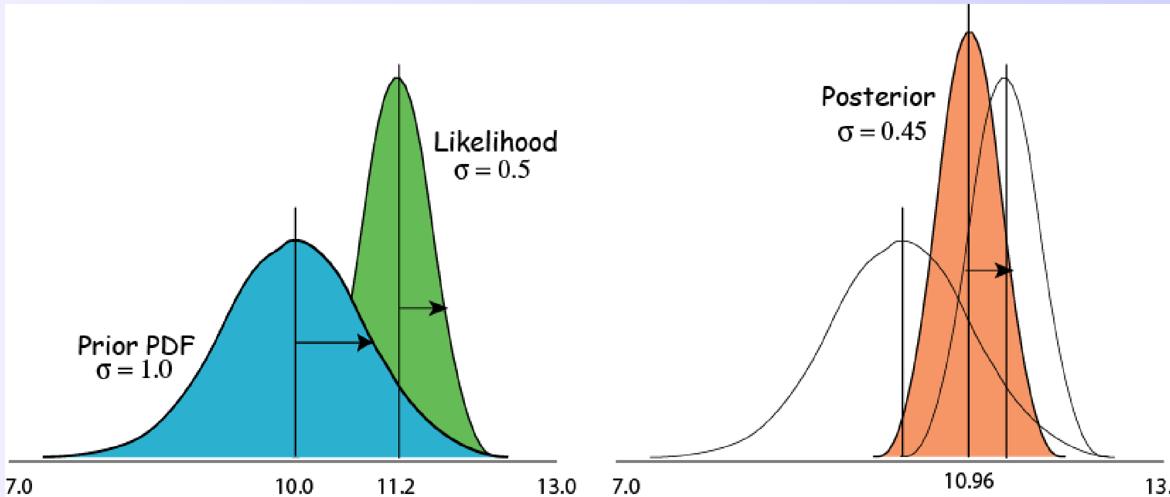
Likelihood

Posterior

The posterior PDF becomes a Gaussian centred at the value of 10.96 μ g with standard deviation σ = $(1/5)^{1/2}$ ¼ 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

1702-1761

Bayes' rule (1763)

Conditional PDFs

Assumptions

data

model parameters

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

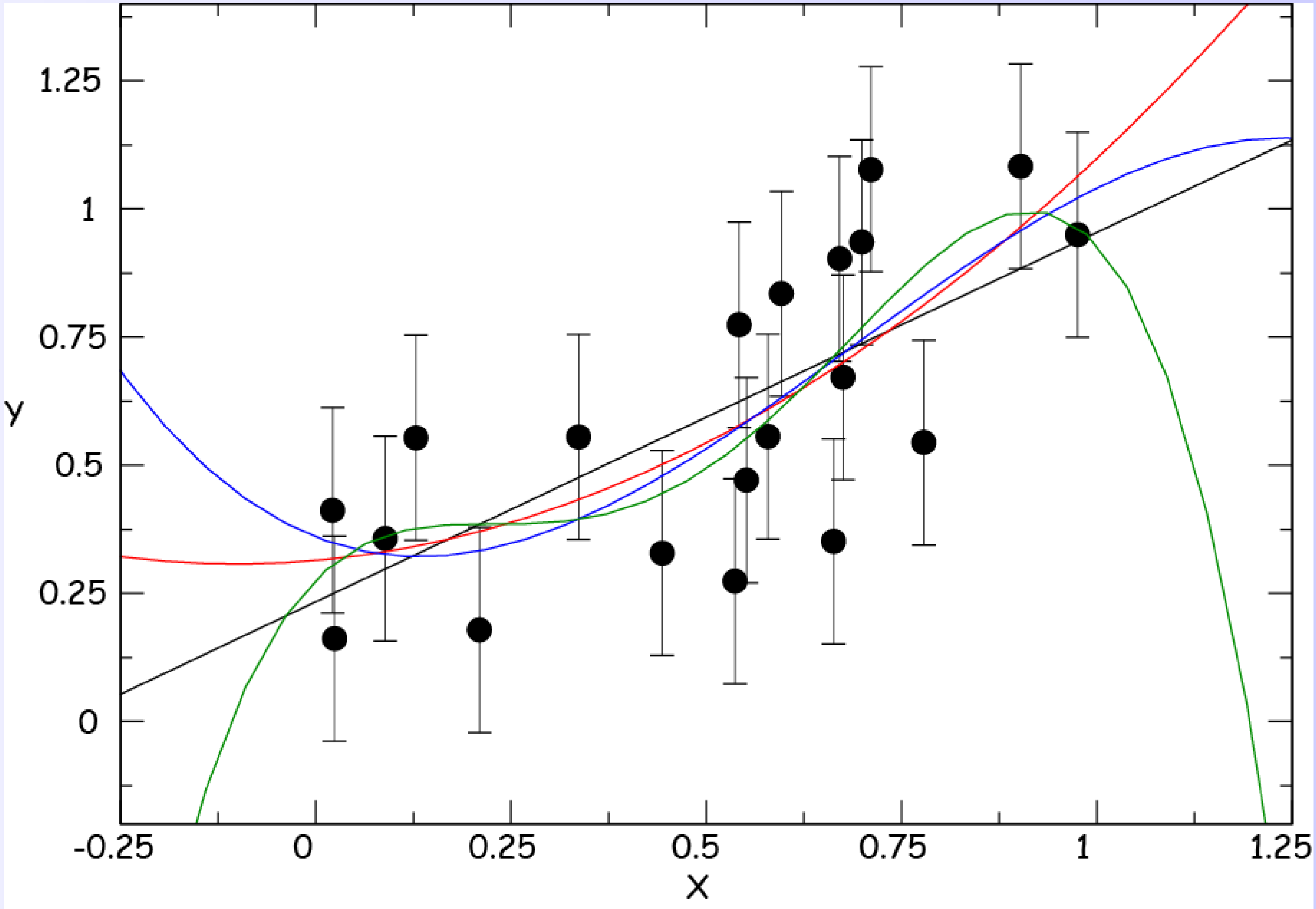
What is known after the data are collected

Measuring fit to data

What is known before the data are collected

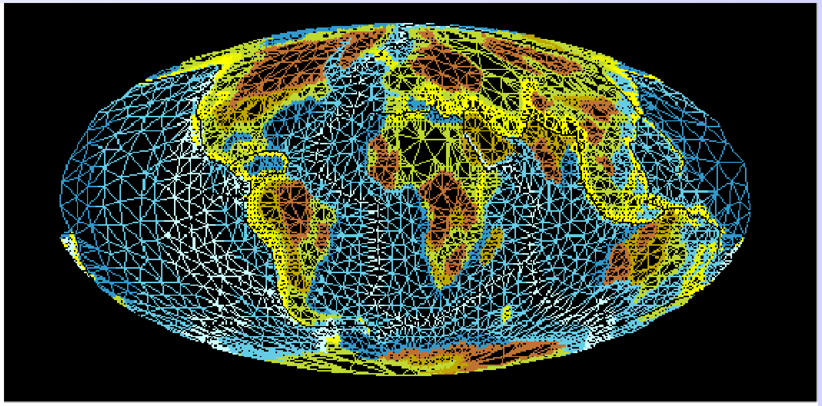
Trans-dimensional sampling: A regression example

Which curve produced that data ?

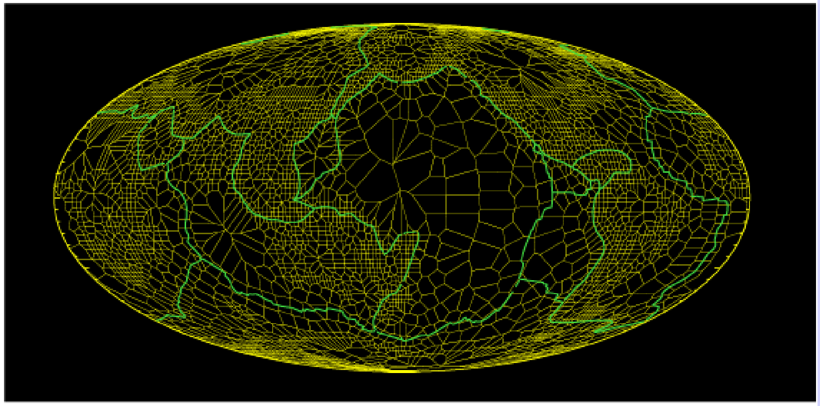


From Sambridge et al. (2006)

Irregular grids in seismology



Delaunay Tessellation



VoronoiTessellation

Gudmundsson & Sambridge (1998)