Bayesian Parameter Inference and Model Comparison

Alan Heavens and Harry Mootoovaloo
Imperial College London

July 25, 2019

DISCNET

Overview

1. Parameter Inference
   - The posterior $p(parameters|data)$
   - Case study: Eddington 1919 Eclipse expedition
2. Model Comparison
3. Bayesian Hierarchical Models (BHM)
   - Case study: straight line fitting with errors in x and y
   - Really big BHMs

Bayesian Inference

What questions do we want to answer?

Model Comparison:

- Do data support General Relativity or Newtonian gravity?

Notation

- **Data** d; **Model** M; Model **parameters** θ
- Rule 1: write down what you want to know
- Usually, it is the probability distribution for the parameters, given the data, and assuming a model.
- $p(\theta|\mathrm{d}, M)$
- This is the **Posterior**
- To compute it, we use Bayes theorem:

$$p(\theta|d, M) = \frac{p(d|\theta, M)p(\theta|M)}{p(d|M)}$$

- where the **Likelihood** is $\mathcal{L}(\mathrm{d}|\theta) = p(\mathrm{d}|\theta, M)$
- and the **Prior** is $\pi(\theta) = p(\theta|M)$
- $p(\mathrm{d}|M)$ is the **Bayesian Evidence**, which is important for Model Comparison, but not for Parameter Inference.
- Dropping the M dependence

$$p(\theta|d) = \frac{\mathcal{L}(d|\theta)\pi(\theta)}{p(d)}$$

The Posterior

$$p(\theta|d, M)$$

If you just try long enough and hard enough, you can always manage to boot yourself in the posterior. A.J. Liebling.

It is all probability

The Posterior

Everything is focussed on getting at $p(\theta|\mathrm{d})$.

Computing the posterior

$p(\theta|\mathrm{d}) \propto \mathcal{L}(\theta)\,\pi(\theta)$.

We need to analyse the problem:

What are the data, d?
What is the model for the data?
What are the model parameters?
What is the likelihood function $\mathcal{L}(\theta)$?
What is the prior $\pi(\theta)$?

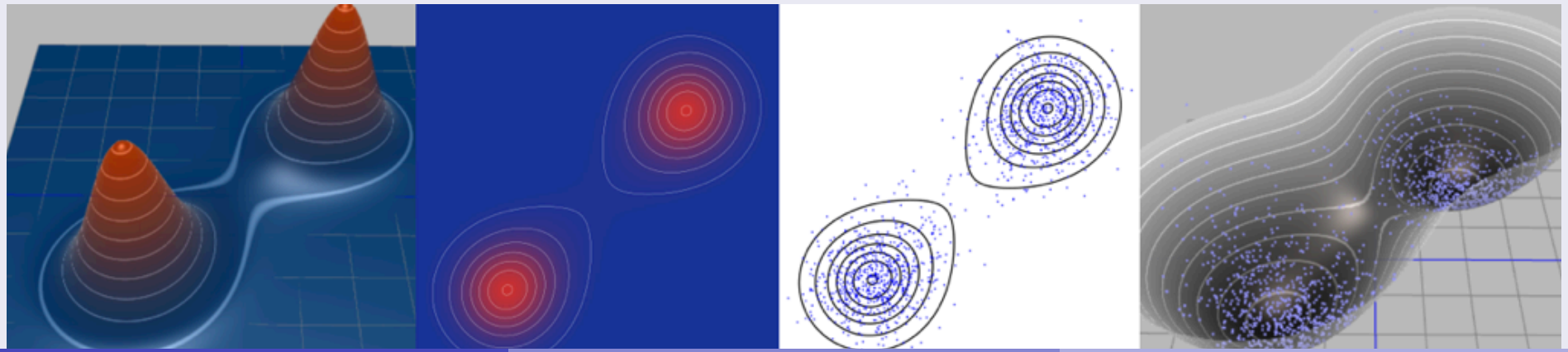
Sampling

The posterior is rarely an analytic function, and evaluating it on a grid in parameter space is usually prohibitively expensive if there are more than 2 or 3 parameters.

MCMC

Standard technique is MCMC (Markov Chain Monte Carlo), where random steps are taken in parameter space, according to a proposal distribution, and accepted or rejected according to the Metropolis-Hastings algorithm. This gives a chain of samples of the posterior (or the likelihood), with an expected number density proportional to the posterior.

MCMC example



Case study: Eddington 1919 Eclipse expedition

In General Relativity, light is bent by the Sun through an angle $\frac{4GM}{rc^2}$.

In Newtonian theory, the bend angle is $\frac{2GM}{rc^2}$.

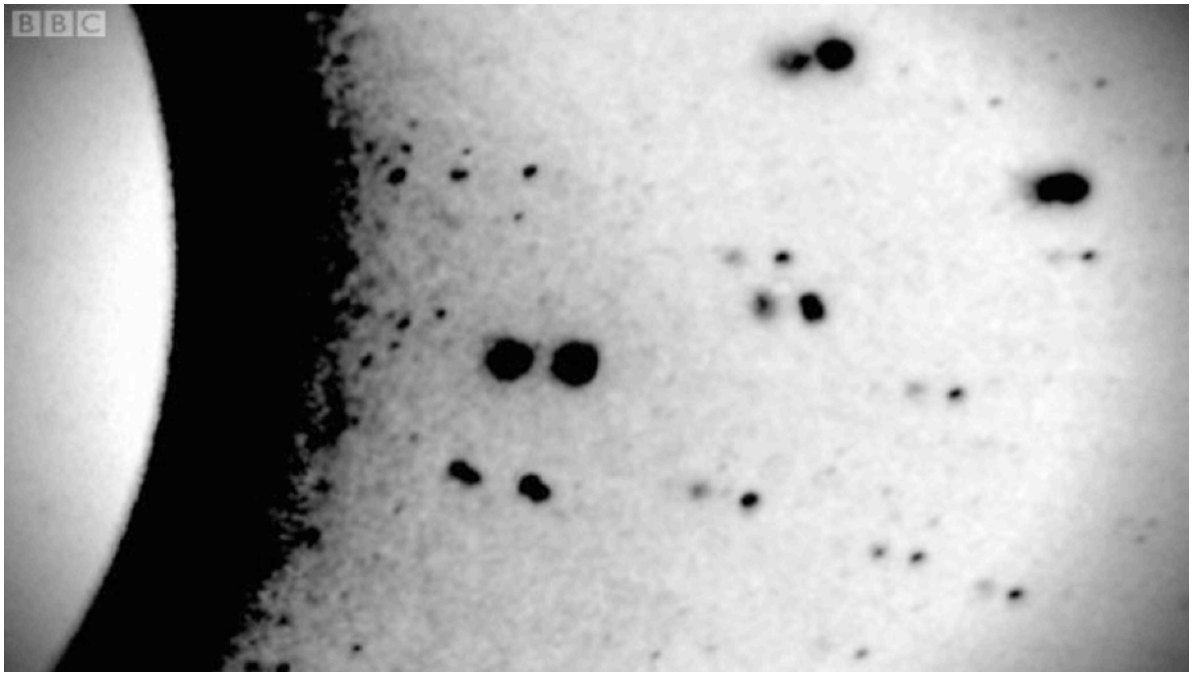


Figure: Illustration of the lensing effect (highly magnified).

If we treat this as a parameter inference problem: bend angle at the limb of the Sun $= \alpha$, and we will infer α.

Case study: Eddington 1919 Eclipse expedition

Analyse the experiment:

- **What are the data**?
- Measurements of displacements (Dx, Dy) of (7) stars, between eclipse plate(s) and a reference plate.
- **What is the model?**
- Displacements are radial, with magnitude α (arcsec) for light grazing the Sun.
- **What are the model parameters?**
- α
- **What is the likelihood function?**
- Measurement errors are Gaussian (assumption!)
- **What prior should we choose?**
- Uniform $\pi(\alpha) =$constant.

Case study: Eddington 1919 Eclipse expedition

Hang on.

- Reference plate may not be centred correctly
- Plates may be rotated with respect to each other
- Eclipse plate may have be scaled in size (thermal effects/different instruments)
- Model for the data is

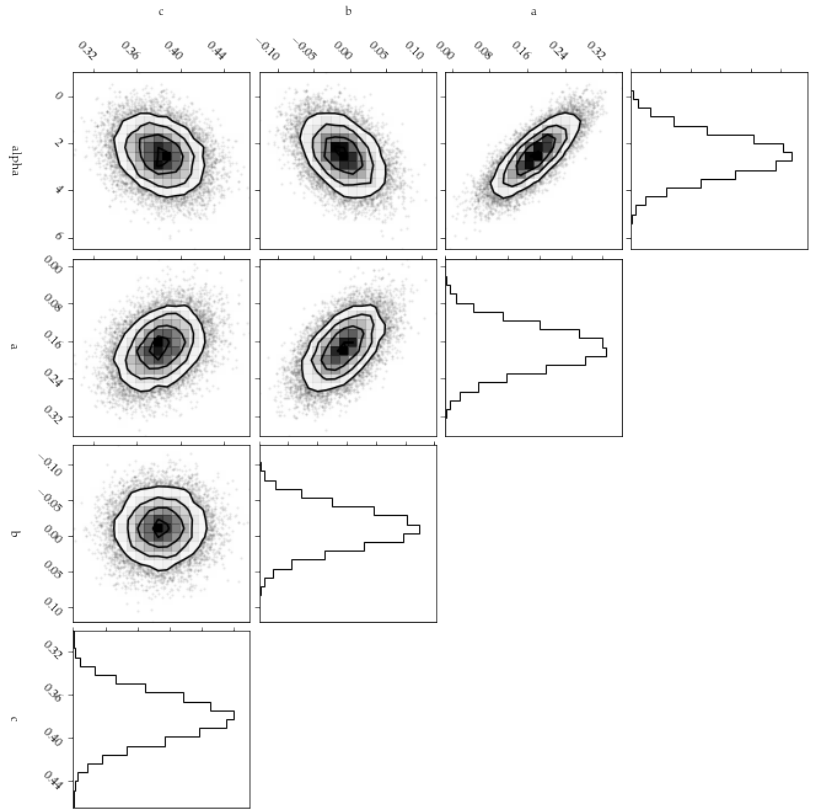
$$\begin{aligned} Dx &= ax + by + c + \alpha E_x \\ Dy &= dx + ey + f + \alpha E_y. \end{aligned} \tag{1}$$

- 7 parameters, including 6 *nuisance parameters*
- Likelihood

$$\mathcal{L} \propto \prod_{\text{stars } i} \exp\left\{ -\frac{[Dx_i - (ax_i + by_i + c + \alpha E_{xi})]^2}{2\sigma_i^2} \right\}$$

Results from Plate II

Using only displacements in R.A. (4 parameters)



Marginalise over nuisance parameters!

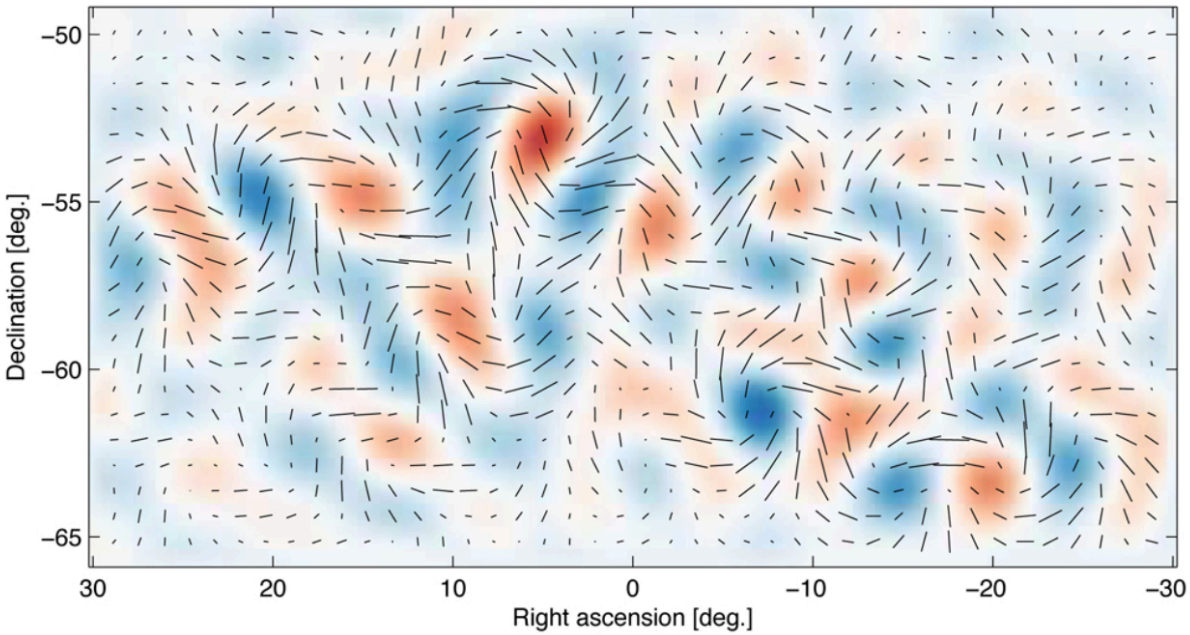
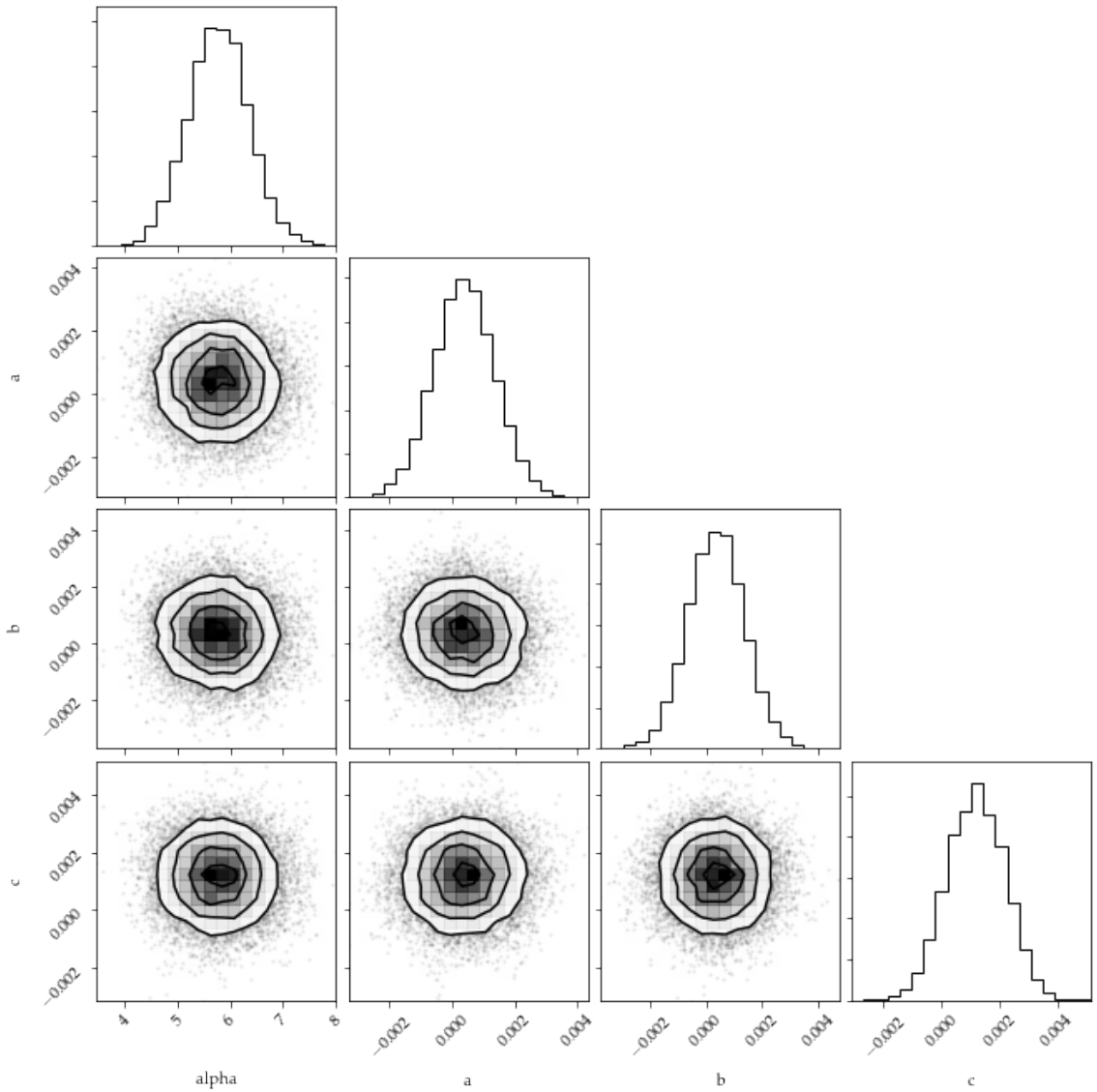


Figure: Exquisite BICEP B-mode CMB map (Credit: BICEP team).

Nuisance Parameters

What if we do not marginalise over a, b, c?



Model Comparison

- Higher-level question than parameter inference
- Which theoretical framework ('model') is preferred, given the data (regardless of the parameter values)
- The models may be completely different (e.g. compare Big Bang with Steady State, to use an old example),
- or variants of the same idea - e.g. add a new parameter
- The sort of question asked here is often 'Do the data favour a more complex model?'
- Clearly in the latter type of comparison the likelihood itself will be of no use - it will always increase if we allow more freedom.
- We compare the **Bayesian Evidence** for each model,

$$p(d|M) = \int \mathcal{L}(\theta|M)\,\pi(\theta|M)\,d\theta$$

which may be a high-dimensional integral and not trivial to evaluate.

Bayesian Hierarchical Models, for more complex problems

If you can, this is how to do it

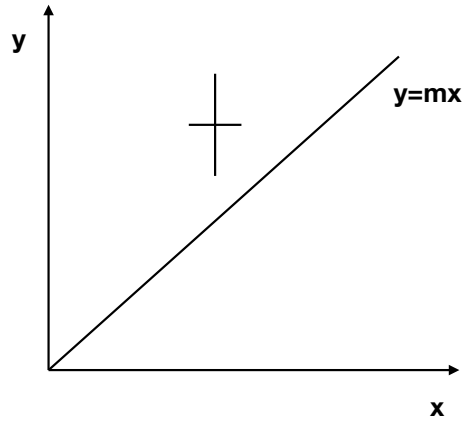
BHM

- We split the inference problem into steps, where the full model is made up of a series of sub-models
- The Bayesian Hierarchical Model (BHM) links the sub-models together, correctly propagating uncertainties in each sub-model from one level to the next.
- At each step ideally we will know the conditional distributions
- The aim is to build a complete model of the data
- Principled way to include systematic errors, selection effects (everything, really)

Case study: straight line fitting

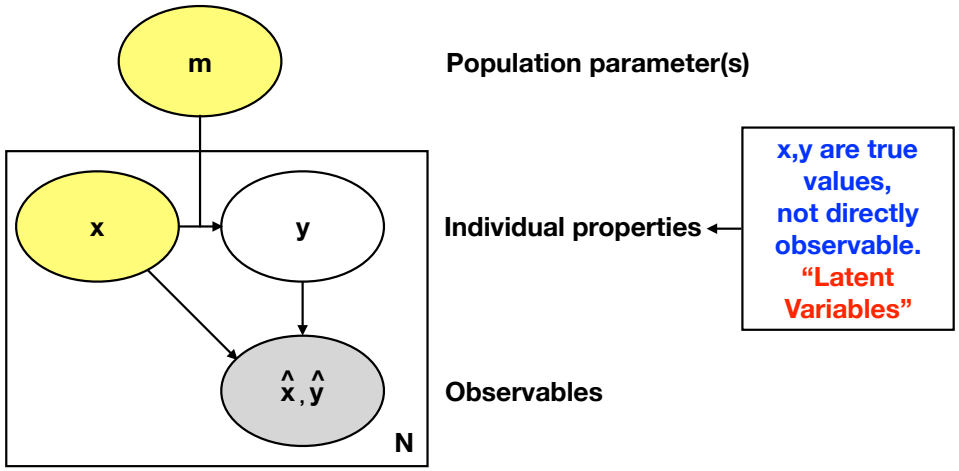
- Let us illustrate with an example. We have a set of **data** pairs $(\hat{x}, \hat{y})$ of noisy measured values of x and y (in fact for simplicity we will have just one pair)
- **Model:** $y = mx$
- **Parameter:** m.
- Complication: $\hat{x}$ and $\hat{y}$ *both* have errors.
- How do we infer m?
- First, apply Rule 1: write down what you want to know.
- It is

$$p(m|\hat{x}, \hat{y})$$



Straight line fitting

How would you forward model it?



- Break problem into two steps.
- There are extra unknowns in this problem (so-called **latent variables**), namely the *unobserved true values* of $\hat{x}$ and $\hat{y}$, which we will call x and y.
- The model connects the *true* variables. i.e.,

$$y = mx.$$

- The latent variables x and y are *nuisance parameters* - we are (probably) not interested in them, so we will marginalise over them.

Hierarchical Bayes vs Ordinary Bayes

- Ordinary Bayes (for given, fixed x):

$$p(m|\hat{y}) \propto p(\hat{y}|m)\, p(m)$$

- Hierarchical Bayes:

$$p(m|\hat{x}, \hat{y}) \propto p(\hat{x}, \hat{y}|m)\, p(m)$$

We do not know the likelihood $p(\hat{x}, \hat{y}|m)$ directly, and we introduce the latent variables:

$$p(m|\hat{x}, \hat{y}) \propto \int p(\hat{x}, \hat{y}, x, y|m)\, p(m)\, dx\, dy$$

Analysis

- Let us now analyse the problem. Manipulating the last equation

$$p(m|\hat{x},\hat{y}) \propto \int p(\hat{x},\hat{y}|x,y,\cancel{m})\, p(x,y|m)\, p(m)\, dx\, dy$$

- $$p(m|\hat{x},\hat{y}) \propto \int p(\hat{x},\hat{y}|x,y)\, p(y|x,m)\, p(x|\cancel{m})\, p(m)\, dx\, dy$$

  This splits the problem into a noise term, a theory term, and priors. We can write all of these down.
- Here, the theory is deterministic:

$$p(y|x,m) = \delta(y - mx)$$

  Integration over y is trivial with the Dirac delta function:

$$p(m|\hat{x},\hat{y}) \propto \int p(\hat{x},\hat{y}|x,mx)\, p(x)\, p(m)\, dx.$$

Choose some priors, and integrate, or sample from the joint distribution of m and x:

$$p(m,x|\hat{x},\hat{y}) \propto p(\hat{x},\hat{y}|x,mx)\, p(x)\, p(m)$$

Gibbs sampling results

- We can sample alternately from m and x, using the conditional distributions, to sample $p(m, x|\hat{x}, \hat{y})$, and marginalise over x in the normal MCMC way by simply ignoring the values of x. Here, $\hat{x} = 10$, $\hat{y} = 15$, and both have gaussian errors with unit variance.

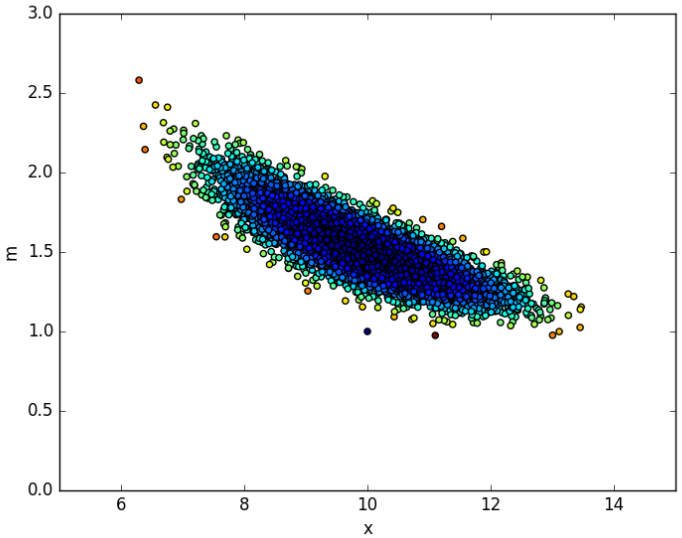


Figure: Gibbs sampling of the latent variable x, and the slope m.

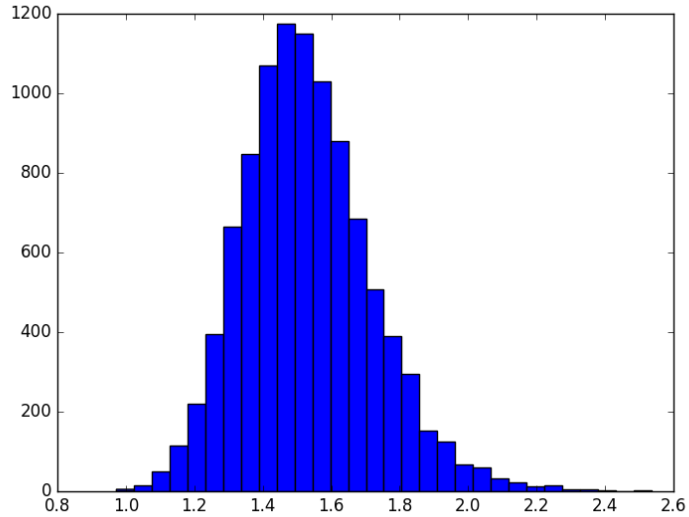


Figure: Gibbs sampling of the slope m.

Question: is this the most probable slope?

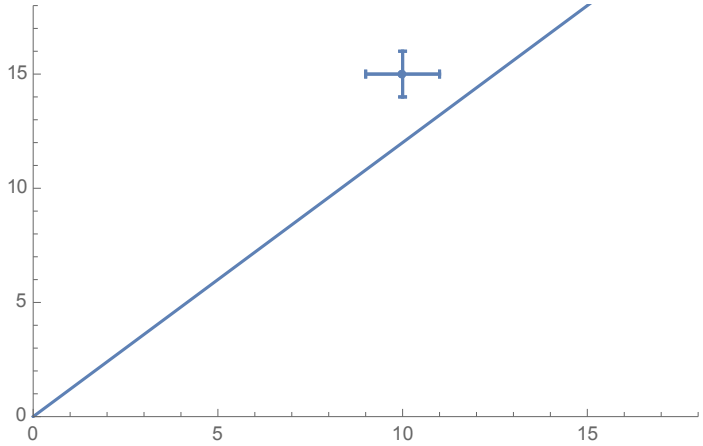


Figure: Noisy data

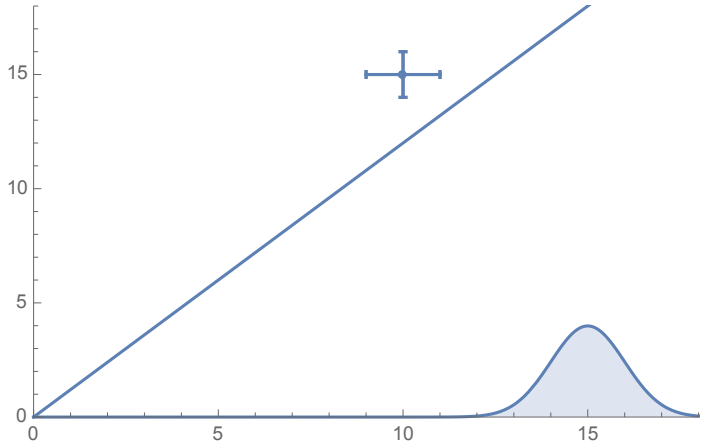


Figure: Yes! - there is a prior on x . . .

Bayesian Hierarchical Models

Computing the posterior

$p(\theta|\mathrm{d})$ may be impossible to calculate directly

e.g. p(cosmology parameters θ|shapes of galaxies d)

Solution: make the problem MUCH harder:

Compute the joint probability of the cosmological parameters *and the shear map*

Joint distribution

$$\mathrm{p}(\theta \,|\, \mathrm{d}) = \int \mathrm{p}(\theta, \mathrm{map} \,|\, \mathrm{d})\, \mathrm{d}(\mathrm{map})$$

$$\mathrm{p}(\theta, \mathrm{map} \,|\, \mathrm{d}) \propto \mathcal{L}(\mathrm{d} \,|\, \theta, \mathrm{map})\, \mathrm{p}(\mathrm{map}|\theta)\, \pi(\theta)$$

Joint map, parameter sampling

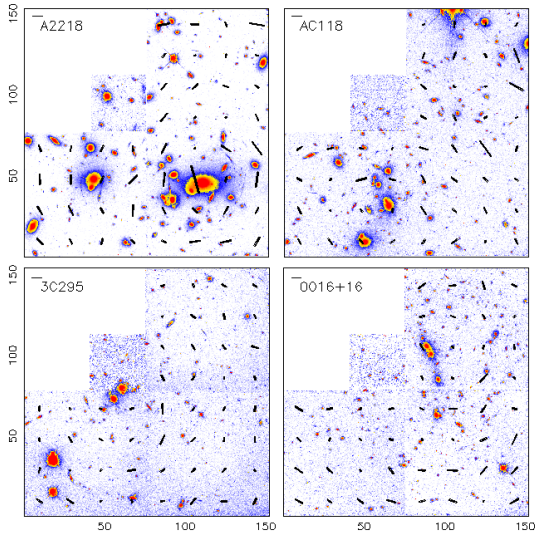


Figure: From Smail et al. 1997.

Latent parameters

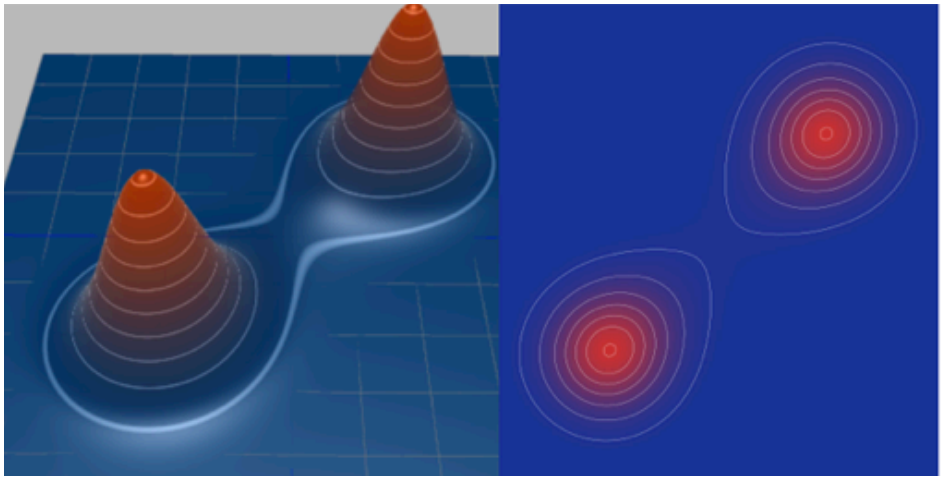
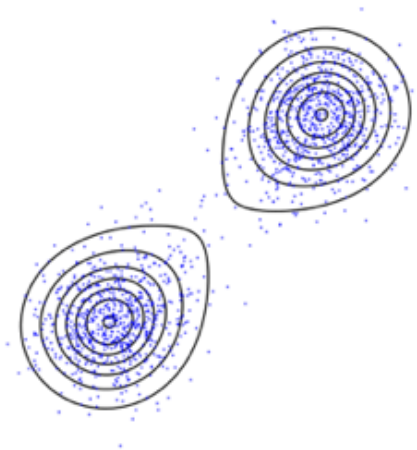
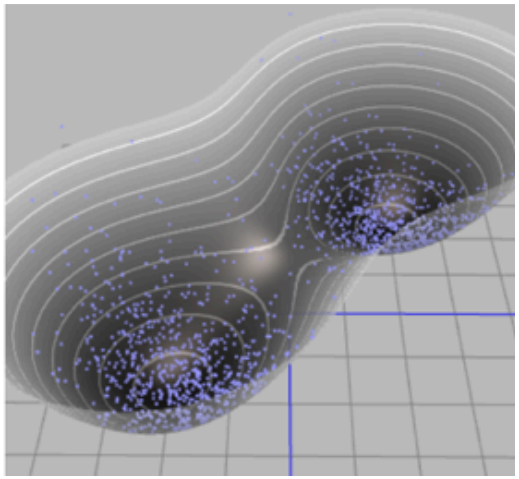
Each pixel in the map is a parameter

10 cosmological parameters, plus 1,000,000 shear values

One million-dimensional probability distribution to calculate...

Sampling in *very* high dimensions

- MCMC: Metropolis-Hastings fails since it is very hard to devise an efficient proposal distribution
- Gibbs sampling: effective if conditional distributions are known
- Hamiltonian Monte Carlo (HMC) works in very high dimensions (e.g. using Stan)

Summary of BHM

- Bayesian Hierarchical Models are a way to build a statistical model of the data by splitting into steps
- Typically, decomposing into steps exposes what is needed - typically many conditional distributions
- For complex data, this may be the *only* viable way to build the statistical model
- The decomposition is usually very natural and logical
- The model allows the proper propagation of errors from one layer to the next,
- including a proper treatment of systematics
- One can often use efficient sampling algorithms to sample from the posterior - precisely what one wants from a Bayesian statistical analysis