Statistical Tools (for Cosmology)

Alan Heavens

Imperial Centre for Inference and Cosmology (ICIC)
Imperial College, London

a.heavens@imperial.ac.uk

YETI meeting: Gravitational Probes of Fundamental Physics
Durham 8-11 January 2017

ICIC

Imperial College London

Overview

The need to do it right

- It's the basis of the scientific method
- Doing the statistics wrong may lead to far-reaching conclusions that may be incorrect (e.g., BICEP)
- The naïve, simple way may be totally misleading, and wrong.

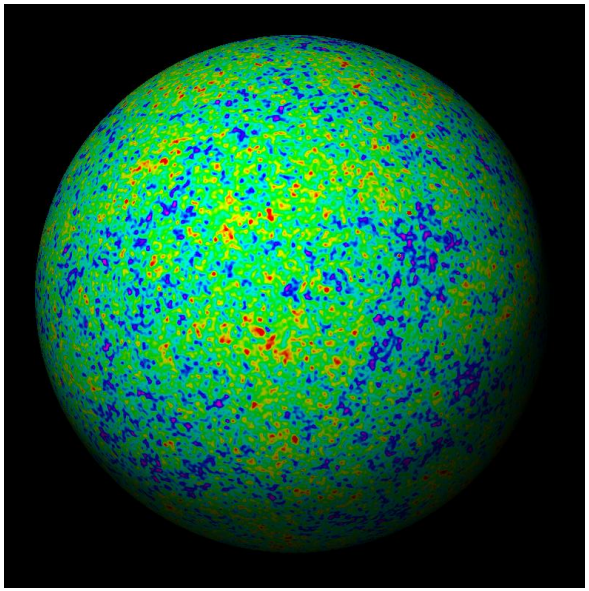


Figure 1: WMAP (credit: M. Tegmark)

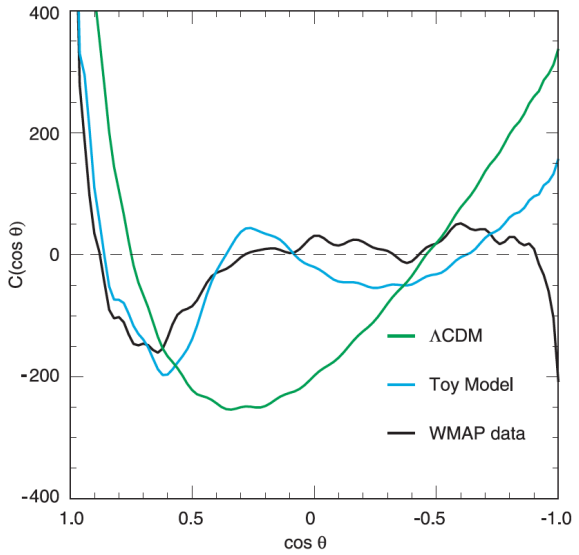


Figure 2: WMAP temperature correlation function (Spergel et al 2003)

Parameter Inference and Bayes' Theorem

- **Rule 1: Write down what you want to know.**
- Given some data $\vec{x}$ and some prior information, what is the probability distribution for the model parameters $\vec{\theta}$? $\boxed{p(\vec{\theta}|\vec{x})}$
- This **posterior distribution** can be written:

$$\boxed{p(\vec{\theta}|\vec{x}) = \frac{p(\vec{x}|\vec{\theta})p(\vec{\theta})}{p(\vec{x})}} \quad \textit{Bayes' Theorem}$$

- $p(\vec{\theta}) =$ **prior** pdf of parameters, often written $\pi(\vec{\theta})$
- $p(\vec{x}|\vec{\theta}) =$ **likelihood** of the data given model parameters. It is treated as an unnormalised function of $\vec{\theta}$
- $p(\vec{\theta}|\vec{x}) =$ **posterior** probability of the parameters, normalised by
- $p(\vec{x}) =$ **evidence**
- All probabilities are implicitly conditional on the model M, and treat probability as a **degree of belief**.

Sampling distribution

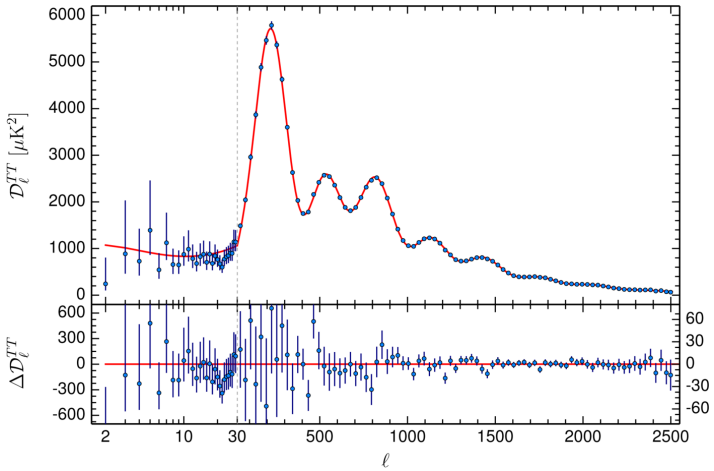


Figure 3: CMB power spectrum (Planck Collaboration 2015)

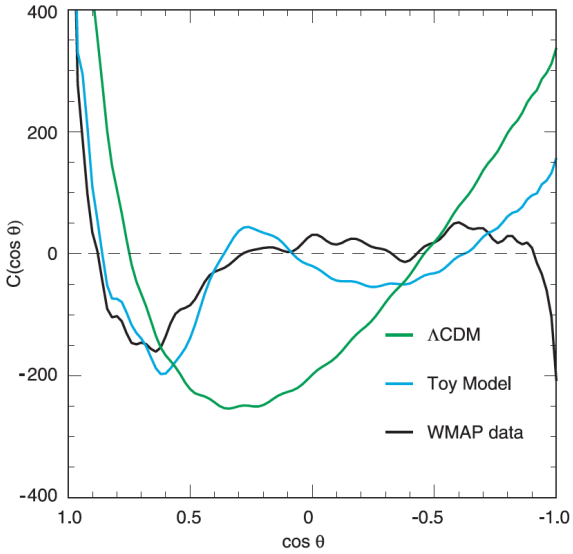


Figure 4: WMAP temperature correlation function (Spergel et al 2003)

We need to know the **sampling distribution** = probability of getting *any* data set, given a model and its parameters.

The Prior

- **Frequentists** don't like priors.
- **Bayesians** embrace them. Interpreting probability as a state of knowledge, then having to specify the prior state (before doing the experiment) makes perfect sense.
- For parameter inference, the prior becomes unimportant as more data are added and the likelihood dominates, but cosmologists are rarely in this luxurious position.
- Generally we want uninformative priors if we don't know anything. Subtle problem.
- Common choices are π=constant for location parameters (e.g. mean)
- $\pi \propto 1/\theta$ ('the' Jeffreys Prior) for scale parameters (which must be positive, e.g. variance)
- The posterior from one experiment can be used as a prior for the next experiment (very useful for combining experimental results)

Marginalisation

The posterior probability of (say) two parameters is given by **marginalising** (integrating) over the others:

$$p(\theta_i, \theta_j|\vec{x}) = \int_{k\neq i \text{ or } j} d\vec{\theta}_k \, p(\vec{\theta}|\vec{x})$$

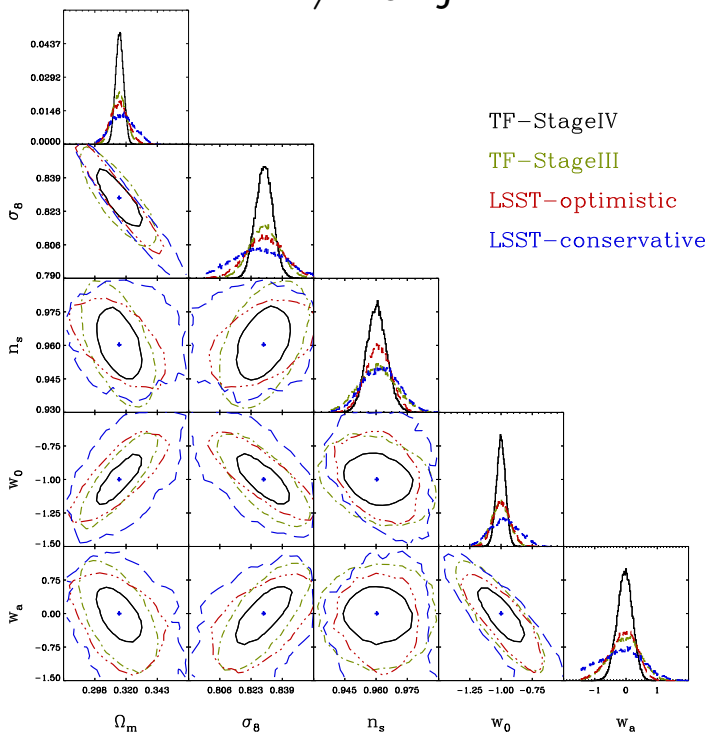


Figure 5: Posterior probabilities for parameters in pairs, marginalised over all others. From Huff et al (2013).

Case Study. BPZ: Bayesian Photometric Redshifts

We follow Benitez (2000), ApJ, 536, 571

- Obtain a posterior for the redshift of a galaxy given measurements of fluxes in some broadband filters (typically 5).

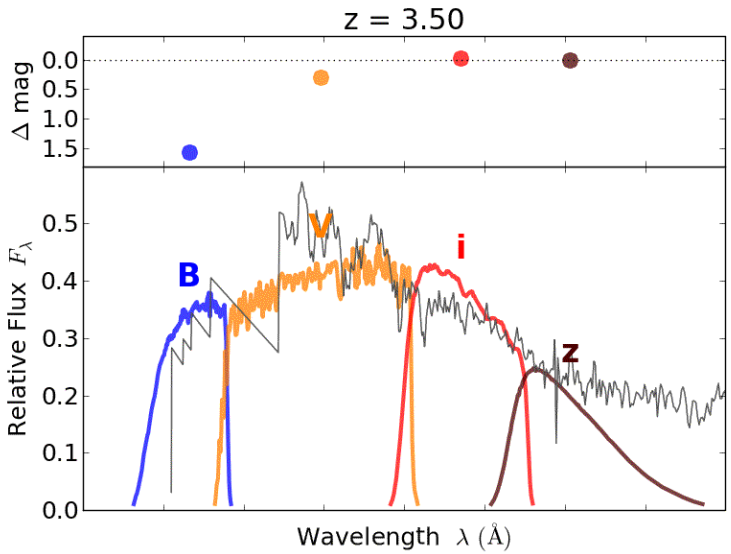


Figure 6: Spectrum and broad band fluxes

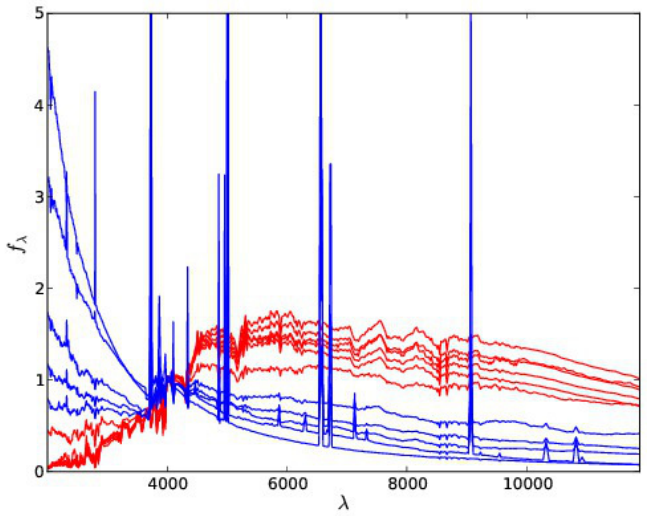


Figure 7: Template spectra

BPZ: likelihood and posterior

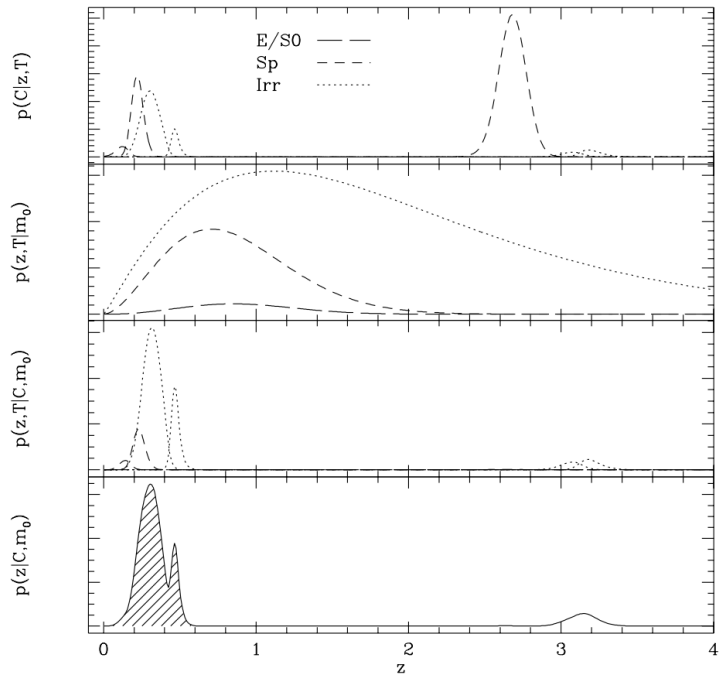


Figure 8: From Benitez (2000)

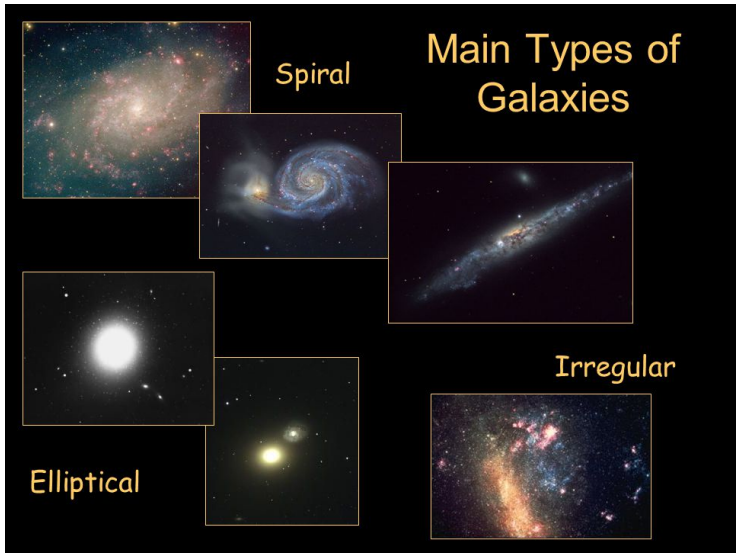


Figure 9: Galaxy Types (credit: tes.com)

- Sometimes the likelihood or posterior can be characterised by a mean and a variance. Not here.
- Marginalising over the template type gives a rich posterior that has no obvious frequentist analogue

Profile likelihoods

Profile likelihoods (where the likelihood is maximised wrt some parameters) makes little sense from a Bayesian perspective.

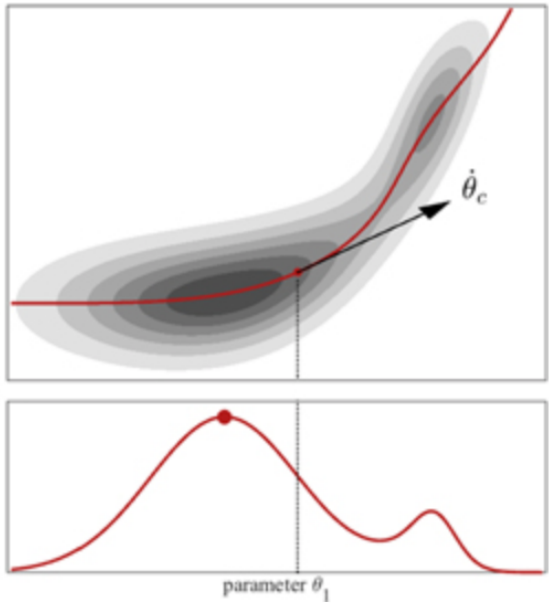


Figure 10: Profile likelihood. From Boiger et al (2016).

Model Comparison

- **A higher-level question** than parameter inference, in which one wants to know which theoretical framework ('model') is preferred, given the data (regardless of the parameter values)
- The models may be completely different, e.g. **General Relativity vs MOND vs** $f(R)$,
- or variants of the same idea. e.g. flat Universe vs. model with curvature
- The sort of question asked here is essentially 'Do the data favour a more complex model?'
- The likelihood itself is not the whole story - it will always increase if we allow more freedom.

Bayesian Evidence or Marginal Likelihood

- Rule 1: Write down what you want to know.
- $\boxed{p(M|\vec{x})}$ - the probability of the model, given the data.
- Use Bayes' theorem:

$$p(M|\vec{x}) = \frac{p(\vec{x}|M)\pi(M)}{p(\vec{x})}$$

- $p(\vec{x}|M)$ is the **Bayesian Evidence**, or **Marginal Likelihood**, and is the denominator in Bayes' theorem for parameter inference:

$$p(\vec{\theta}|\vec{x}, M) = \frac{p(\vec{x}|\vec{\theta}, M)\pi(\vec{\theta}|M)}{p(\vec{x}|M)}$$

  where we have written the dependence on the model M explicitly.
- It normalises the posterior (so that it integrates to unity):

$$p(\vec{x}|M) = \int d\vec{\theta}\, p(\vec{x}|\vec{\theta}, M)\pi(\vec{\theta}|M).$$

Bayesian Evidence

- The relative probabilities of two models is then

$$\frac{p(M'|\vec{x})}{p(M|\vec{x})} = \frac{\int d\vec{\theta'}\, p(\vec{x}|\vec{\theta'}, M')\pi(\vec{\theta'}|M')}{\int d\vec{\theta}\, p(\vec{x}|\vec{\theta}, M)\pi(\vec{\theta}|M)} \times \frac{\pi(M')}{\pi(M)}$$

- The first ratio is the **Bayes Factor**,

$$B \equiv \frac{\int d\vec{\theta'}\, p(\vec{x}|\vec{\theta'}, M')\, \pi(\vec{\theta'}|M')}{\int d\vec{\theta}\, p(\vec{x}|\vec{\theta}, M)\, \pi(\vec{\theta}|M)}.$$

Model Comparison

$$B \equiv \frac{\int d\vec{\theta}'\, p(\vec{x}|\vec{\theta}', M')\, \pi(\vec{\theta}'|M')}{\int d\vec{\theta}\, p(\vec{x}|\vec{\theta}, M)\, \pi(\vec{\theta}|M)}$$

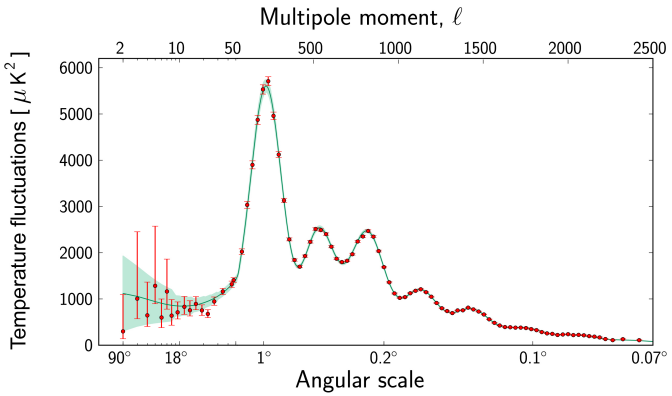


Figure 11: Planck power spectrum, and LCDM model with most probable parameters. Models which cannot reproduce the curve, or can only if the parameters are fine-tuned, will be disfavoured. Credit: Planck

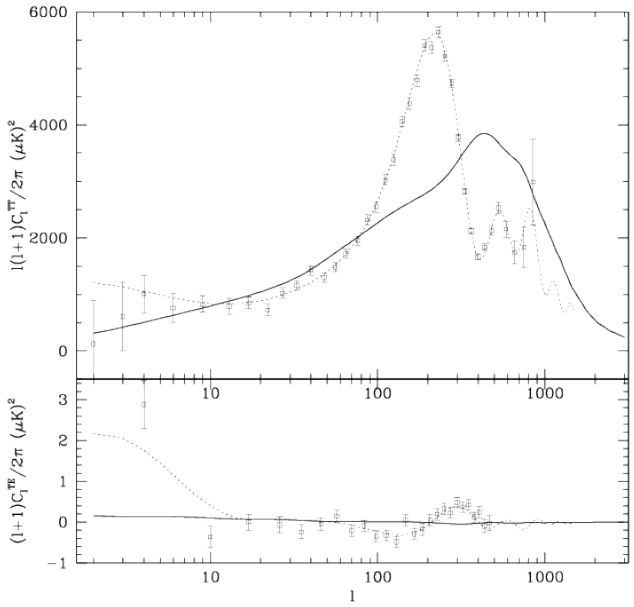


Figure 12: Cosmic String model predictions for CMB (Wyman et al 2005)

Bayesian evidence

Challenges: The evidence requires a multidimensional integration over the likelihood and prior, and this may be *very* expensive to compute.

- **Fisher matrix approach:** assume the likelihood is a multivariate gaussian (Laplace approximation)
- **Approximations:** e.g., AIC and BIC may be unreliable as they are based on the best-fit χ^2, and from a Bayesian perspective we want to know how much parameter space would give the data with high probability. Also don't include the prior. Not Bayesian.
- **Nested sampling** (e.g., multinest, polychord, diffusive nested sampling), where one tries to sample the likelihood in an efficient way. State-of-the-art.

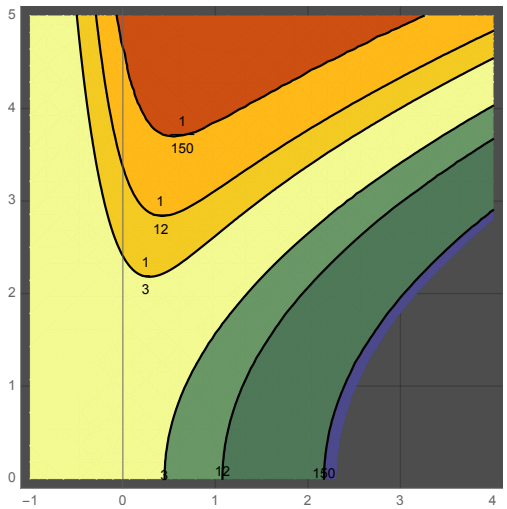
Gaussian Example

Let M_0 be $x \sim \mathcal{N}(0, \sigma^2)$, and M_1 be $x \sim \mathcal{N}(\mu, \sigma^2)$, where the prior on μ is gaussian with variance Σ^2. Let the measurement be $x = \lambda\sigma$.

$$B_{01} = \sqrt{1 + \frac{\Sigma^2}{\sigma^2}} \exp\left[-\frac{\lambda^2}{2(1 + \frac{\sigma^2}{\Sigma^2})}\right]$$

If $\lambda \gg 1$, then B_{01} can be $\ll 1$ and M_1 is favoured. If $\lambda \simeq 1$ and $\sigma \ll \Sigma$, then M_0 is favoured (Occam's razor). If likelihood is much broader than prior, $\sigma \gg \Sigma$ then $B_{01} \simeq 1$ and nothing has been learned.

Figure 13: $x = \log_{10}(\Sigma/\sigma)$; $y = \mathrm{datum}/\sigma$.



Bayesian Hierarchical Models

- Complex data analysis problems can often be split into steps: 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 we will need to know conditional distributions.

Analytic Example: straight line fitting

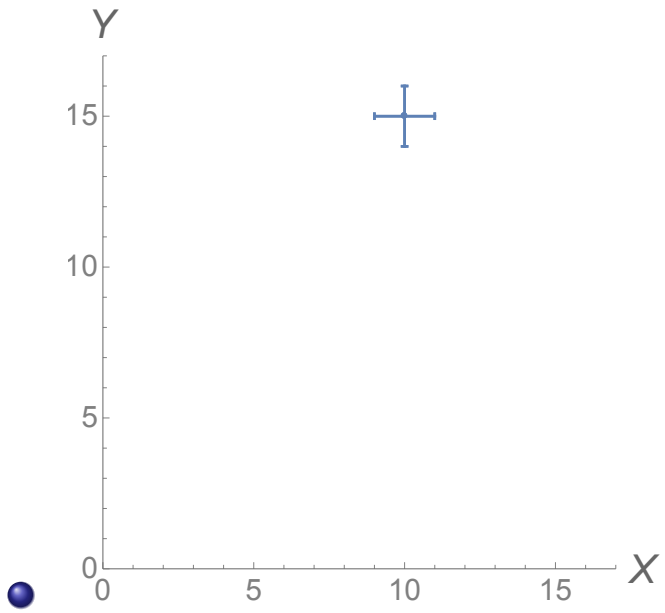


Figure 14: Errors in both variables

- Data: (X, Y)
- Model: $y = mx$
- Parameter (to be inferred): m.
- Complication: X and Y both have errors.
- Rule 1: write down what you want to know.
- $\boxed{p(m|X, Y)}$

Straight line fitting

- Break problem into steps.
- There are extra unknowns in this problem (so-called **latent variables**), namely the unobserved true values of X and Y, which we will call x and y.
- The model connects the *true* variables.
- $y = mx$
- The latent variables x and y are **nuisance parameters** - we are (probably) not interested in them, so we will end up marginalising over them.
- Introducing these latent variables is **Data Augmentation**

Analysis

- Bayes:

$$p(m|X,Y) \propto p(X,Y|m)\, p(m)$$

  Let us assume $p(m)$=constant.

- Introduce the latent variables x, y, and marginalise over them:

$$p(m|X,Y) \propto \int p(X,Y,x,y|m)\, dx\, dy$$

- Manipulate:

$$p(m|X,Y) \propto \int p(X,Y|x,y,m)\, p(x,y|m)\, dx\, dy$$

- $p(X,Y|x,y,m) = p(X,Y|x,y)$ (errors do not depend on m)
- $p(x,y|m) = p(y|x,m)p(x|m)$
- $p(y|x,m) = \delta(y - mx)$ (model is deterministic)

Analysis

- Integration over y is trivial with the Dirac delta function:

$$p(m|X,Y) \propto \int p(X,Y|x,mx)\, p(x)\, dx$$

  Prior on x is independent of m, so we have written $p(x|m) = p(x)$.

- Assume errors in X and Y are independent Gaussians with unit variance, and take a uniform prior for x:

- $$p(m|X,Y) \propto \int e^{-\frac{1}{2}(X-x)^2} e^{-\frac{1}{2}(Y-mx)^2}\, dx$$

- Complete the square and integrate

$$\boxed{p(m|X,Y) \propto \frac{1}{\sqrt{1+m^2}}\, e^{-\frac{(-mX+Y)^2}{2(1+m^2)}}}$$

Results

We have marginalised analytically over x, but we can also investigate the joint distribution of x and m:

$$p(x, m|X, Y) \propto e^{-\frac{1}{2}(X-x)^2} e^{-\frac{1}{2}(Y-mx)^2}.$$

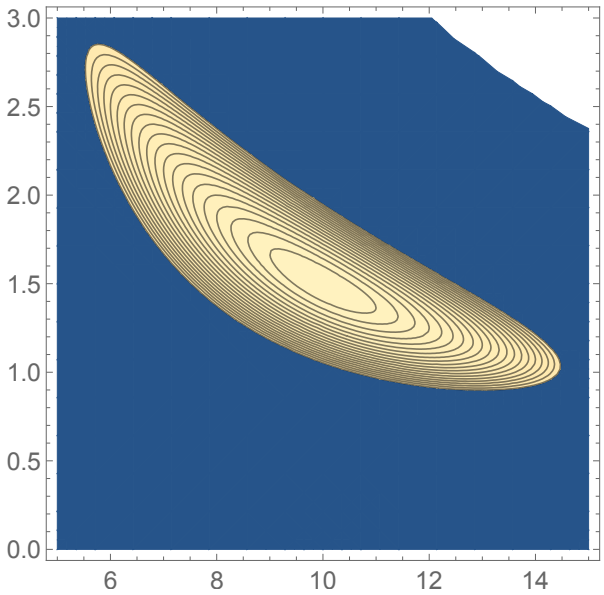


Figure 15: Posterior distribution of x and m.

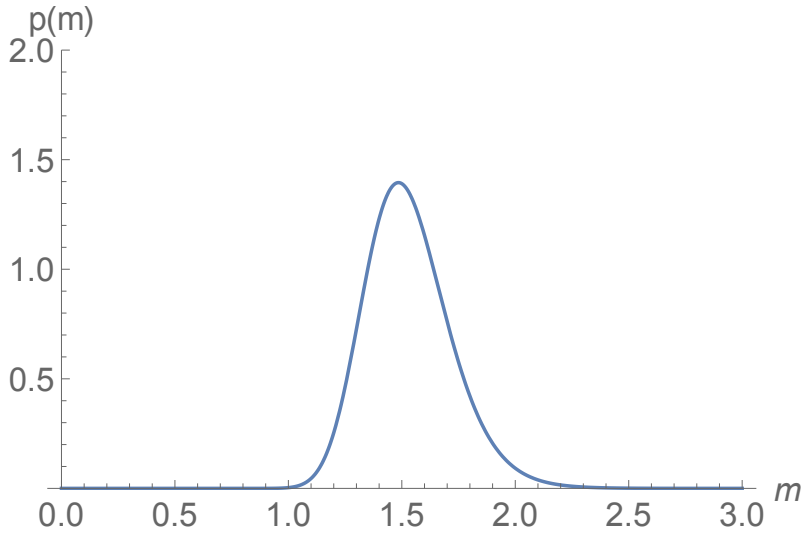


Figure 16: Posterior distribution of the slope m, for $X = 10$, $Y = 15$.

Sampling the posterior (or likelihood)

Probabilities are rarely analytic functions. We can evaluate them on a grid in parameter space, but this is hopeless in many dimensions.

- Instead, we **sample** the parameter space, with an expected number density $n(\vec{\theta})$ proportional to the *target density* (e.g. likelihood or posterior).
- The (unnormalised) target density is approximated by a set of delta functions

$$p(\vec{\theta}) \propto n(\vec{\theta}) \simeq \sum_{i=1}^{N} \delta(\vec{\theta} - \vec{\theta}_i)$$

- from which we can estimate any integrals (such as the mean, variance):

$$\left\langle f(\vec{\theta}) \right\rangle \simeq \frac{1}{N} \sum_{i=1}^{N} f(\vec{\theta}_i).$$

- If we sample from the likelihood, and want the posterior, we can weight the points with the prior. This is **Importance Sampling**

Markov Chain Monte Carlo (MCMC)

Markov: Each point in the chain depends only on the previous point.

Metropolis-Hastings algorithm:

- Take a step away from the present point, using a **proposal distribution** $q(\vec{\theta}^*|\vec{\theta})$ = probability of a move from $\vec{\theta}$ to $\vec{\theta}^*$.
- Accept it with a probability which depends on the ratio of the new and old target densities:

$$p(acceptance) = min\left[1, \frac{p(\vec{\theta}^*)q(\vec{\theta}^*|\vec{\theta})}{p(\vec{\theta})q(\vec{\theta}|\vec{\theta}^*)}\right]$$

  If the proposal distribution is symmetric, the algorithm simplifies to the **Metropolis algorithm**.
- If new point is rejected, the previous point is **repeated**.

Considerations

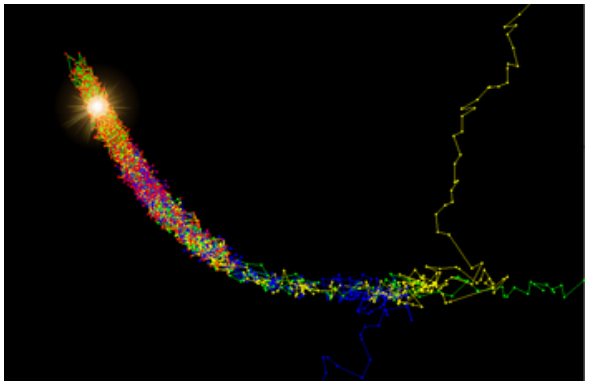


Figure 17: MCMC chain, showing burnin (wikipedia)

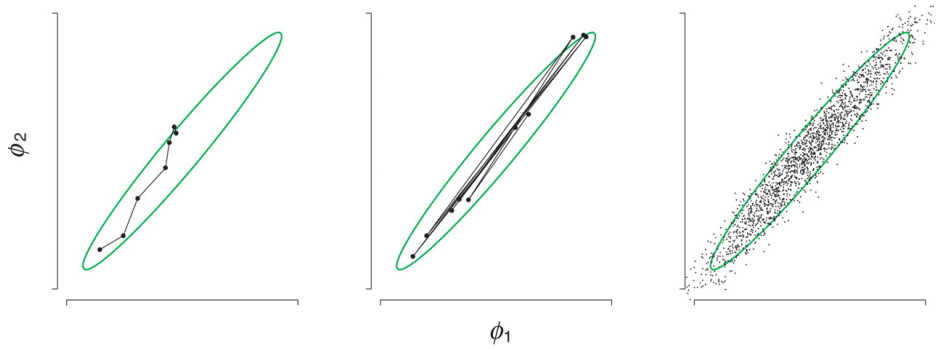


Figure 18: Kuss et al. (2005)

- **Burn-in**. Beginning of the chain is thrown away
- **Proposal distribution** should be neither too small (poor *mixing*- i.e. target is not explored efficiently), nor too large (too many rejections)
- Rule-of-thumb: accept ~ 0.25 of points
- If you change the proposal distribution, you have to start again
- Points will be correlated to some degree. Chain is often *thinned*
- **A convergence test** must be done (typically Gelman-Rubin)

Alternatives to Metropolis-Hastings

Gibbs Sampling: useful if you know the conditional distributions $p(\theta_i|\vec{\theta})$. All points are accepted.

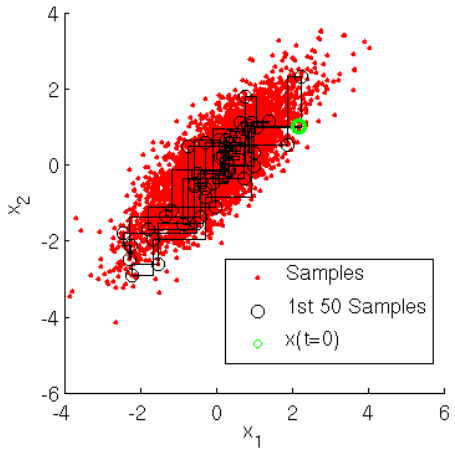


Figure 19: Gibbs sampling (credit: D. Stansbury)

Hamiltonian/Hybrid Monte Carlo (HMC): useful in very high dimensional spaces, where finding an effective proposal distribution is hard. Needs derivatives. See e.g., arXiV:0906.0664 for details of algorithm.

Gibbs sampling for straight line fit

- Exercise: show that the conditional distributions of m given x, and x given m, are

$$p(m|x,X,Y) \sim \mathcal{N}\left(\frac{Y}{x}, \frac{1}{x^2}\right);\ p(x|m,X,Y) \sim \mathcal{N}\left(\frac{X+Ym}{1+m^2}, \frac{1}{1+m^2}\right)$$

- We sample alternately from m and x
- **Marginalising over x is trivial:** simply ignore the values of x in the chain.

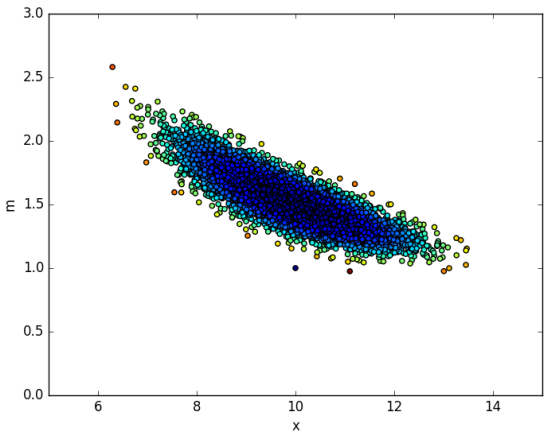


Figure 20: Gibbs sampling of the latent variable x, and the slope m.

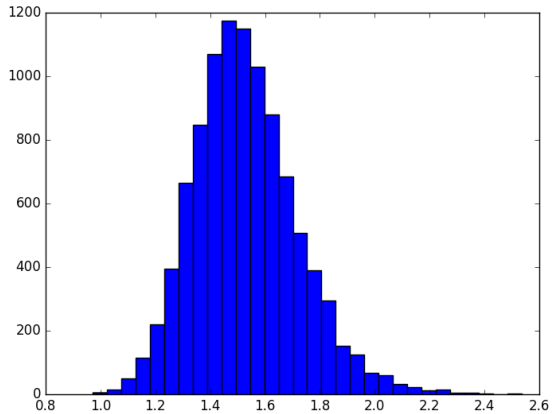


Figure 21: Gibbs sampling of the slope m.

Complex BHM with Gibbs Sampling

Weak lensing in the CFHTLenS survey (Alsing, Heavens, Jaffe, 2016).

$\sim 500,000$ latent variables

Gibbs sampling.

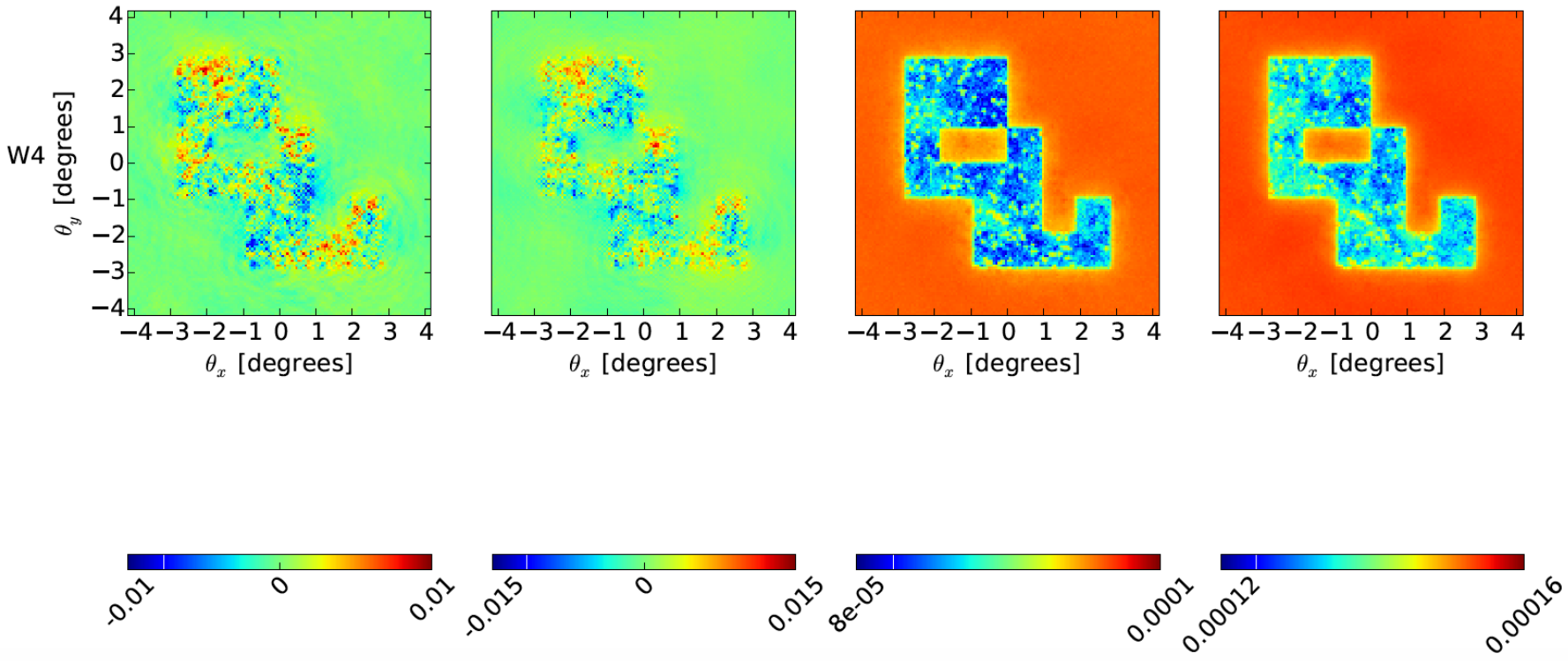


Figure 22: Mean κ (projected mass) map for one CFHTLenS field, and two redshift ranges.

Summary

- Most cosmological analysis is now Bayesian
- Parameter inference is routine, typically using MCMC methods
- Model comparison is increasingly possible
- Complete statistical models of data are needed for principled analysis, and Bayesian Hierarchical Models lead the way
- Very high dimensional inference can be done with Gibbs or Hamiltonian Monte Carlo
- Procedure:
  - What are the data?
  - What is/are the theoretical model(s)?
  - What are the parameters of the model(s)?
  - What is the likelihood function?
  - Apply Rule 1: what do you want to learn?
  - Calculate!

Further Reading

- Data Analysis: a Bayesian Tutorial (Devinder Sivia and John Skilling, CUP)
- Bayesian Methods in Cosmology (Roberto Trotta, https://arxiv.org/abs/1701.01467
- Bayesian Data Analysis (Andrew Gelman et al., CRC Press)
- Information Theory, Inference and Learning Algorithms (David Mackay, CUP)
- Berkeley course on Bayesian Modeling and Inference (Michael I. Jordan).
  `http://www.cs.berkeley.edu/~jordan/courses/260-spring10/lectures/`

ICIC

Imperial College London