Bayesian Hierarchical Models

Alan Heavens

September 6, 2018

ICIC Data Analysis Workshop

Bayesian Hierarchical Models, for more complex problems

If you can, this is how to do it

BHM

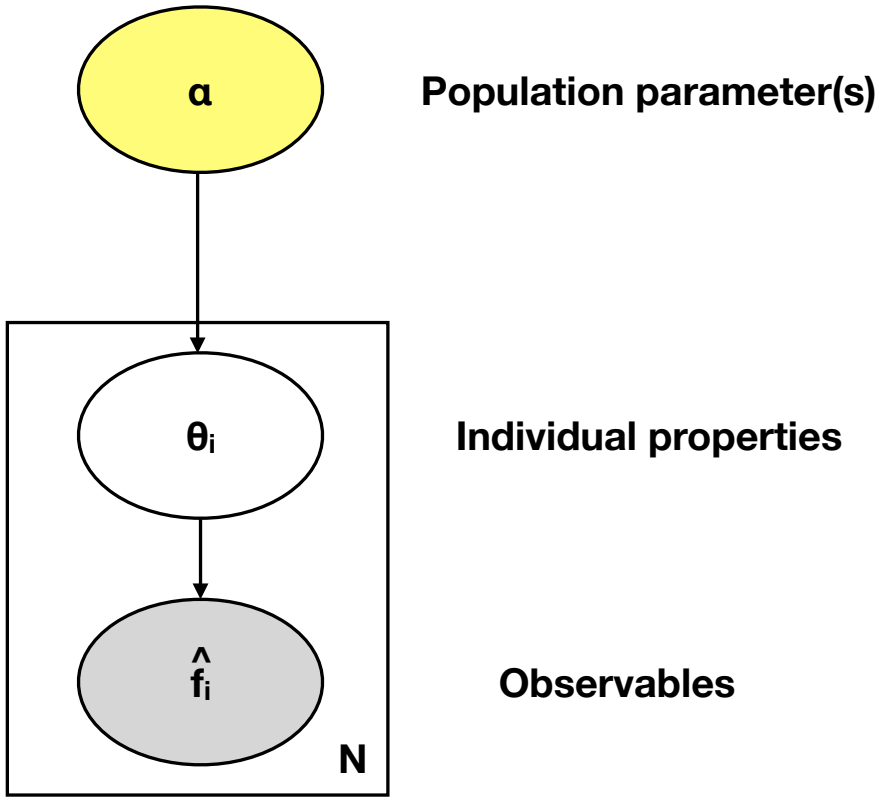
- We split the inference problem into steps, where the full model is made up of a series of sub-models
- The Bayesian Hierarchical Model (BHM) links the sub-models together, correctly propagating uncertainties in each sub-model from one level to the next.
- At each step ideally we will know the conditional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

Bayesian Hierarchical Models

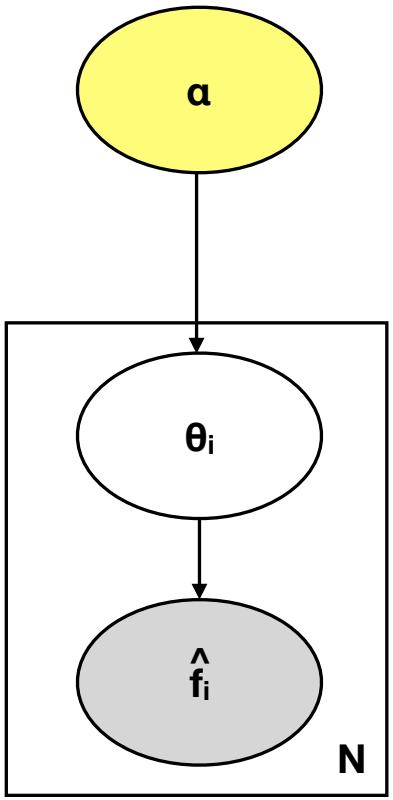
Often used to learn about a *population* from many *individual measurements*. e.g. we measure the number counts of a population of galaxies, but the measured fluxes $\hat{f}_i$ have errors. What are the true number counts?

- Assume (say) a power-law $N \propto f^{-\alpha}$
- Many (unobserved) fluxes θ_i
- Add noise: $\hat{f}_i = \theta_i + n_i$

Number counts



Latent Variables



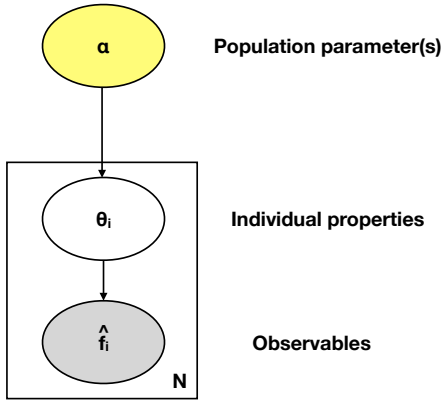
Population parameter(s)

Individual properties

θ are true fluxes, not directly observable. "Latent Variables"

Observables

Ordinary Bayes vs Hierarchical Bayes



- Ordinary Bayes:

$$p(\alpha|\hat{f}) \propto p(\hat{f}|\alpha)\, p(\alpha)$$

- But we do not know $p(\hat{f}|\alpha)$!
- Hierarchical Bayes:

$$p(\theta, \alpha|\hat{f}) \propto p(\hat{f}|\theta, \cancel{\alpha})\, p(\theta, \alpha)$$

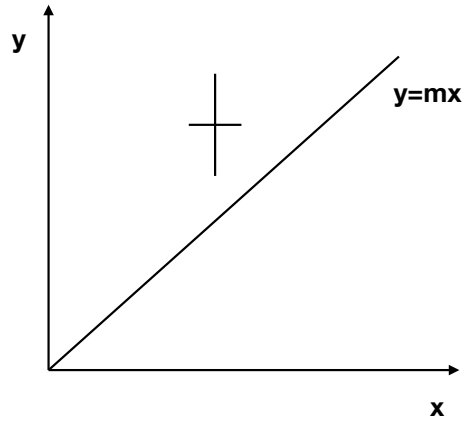
-

$$p(\theta, \alpha|\hat{f}) \propto p(\hat{f}|\theta)\, p(\theta|\alpha)\, p(\alpha)$$

Case study: straight line fitting

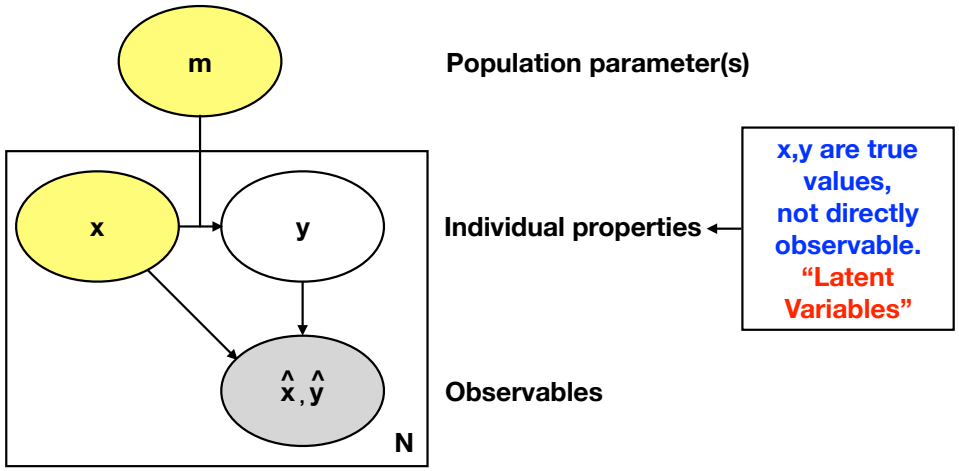
- Let us illustrate with an example. We have a set of **data** pairs $(\hat{x}, \hat{y})$ of noisy measured values of x and y (in fact for simplicity we will have just one pair)
- **Model:** $y = mx$
- **Parameter:** m.
- Complication: $\hat{x}$ and $\hat{y}$ *both* have errors.
- How do we infer m?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

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

- Integrate, or sample from the joint distribution of m and x:

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

Analysis continued

- Repeated from last slide:

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

- Assume errors in x and y are independent Gaussians, and take uniform priors for x and m. For simplicity, let us take $\sigma_{\rm x}^2 = \sigma_{\rm y}^2 = 1$.
-

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

- Complete the square

$$p(m|\hat{x},\hat{y}) \propto \frac{1}{\sqrt{1+m^2}}\, e^{-\frac{(-m\hat{x}+\hat{y})^2}{2(1+m^2)}}.$$

Results

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

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

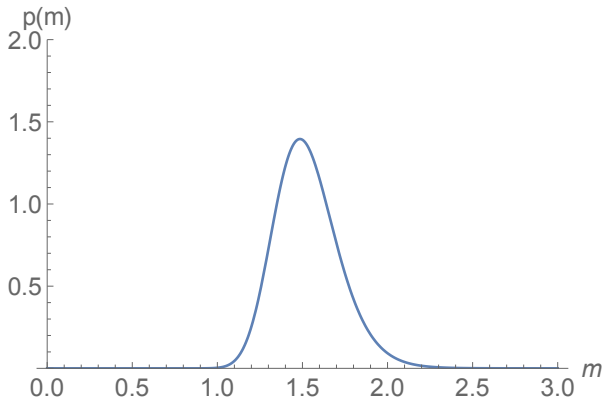


Figure: Unnormalised posterior distribution of the slope m, for $\hat{x} = 10$, $\hat{y} = 15$.

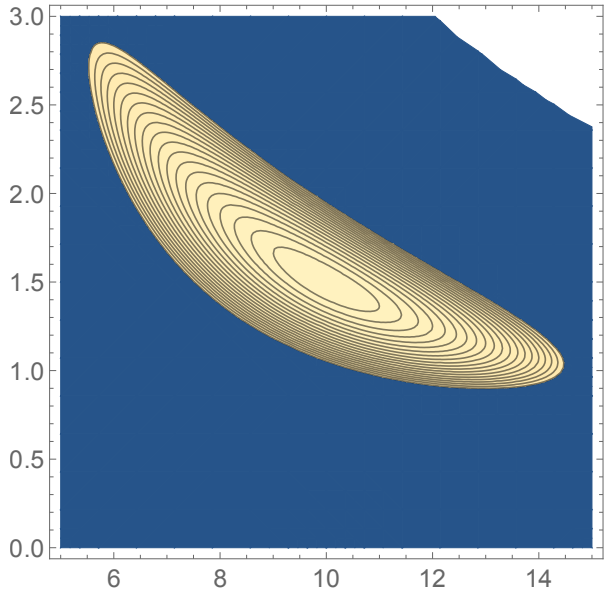


Figure: Unnormalised posterior distribution of the latent variable x, and the slope m.

Gibbs Sampling

Let us see how we would set this up as a Gibbs sampling problem.

- At fixed x, the conditional distribution on m given x is
- $$p(m|\hat{x},\hat{y}) \propto \exp\left[-\frac{(\hat{y}-mx)^2}{2}\right] \propto \exp\left[-\frac{x^2\left(m-\frac{\hat{y}}{x}\right)^2}{2}\right],$$
- i.e.
$$p(m|\hat{x},\hat{y}) \sim \mathcal{N}\left(\frac{\hat{y}}{x},\frac{1}{x^2}\right)$$
is a normal $\mathcal{N}(\mu,\sigma^2)$ distribution (in m).
- The conditional distribution of x given m is
$$p(x|m,\hat{x},\hat{y}) \propto \exp\left[-\frac{(\hat{x}-x)^2}{2}-\frac{(\hat{y}-mx)^2}{2}\right].$$
- After completing the square, this becomes
$$p(x|m,\hat{x},\hat{y}) \sim \mathcal{N}\left(\frac{\hat{x}+\hat{y}m}{1+m^2},\frac{1}{1+m^2}\right)$$

Gibbs results

- Hence 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.

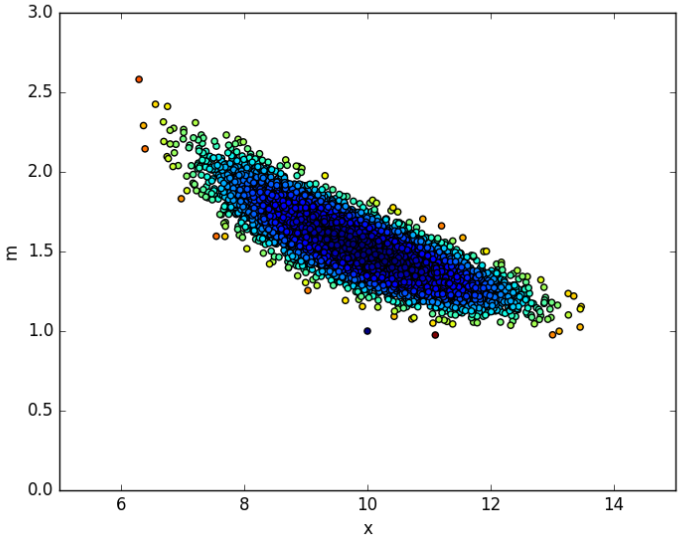


Figure: Gibbs sampling of the latent variable x, and the slope m.

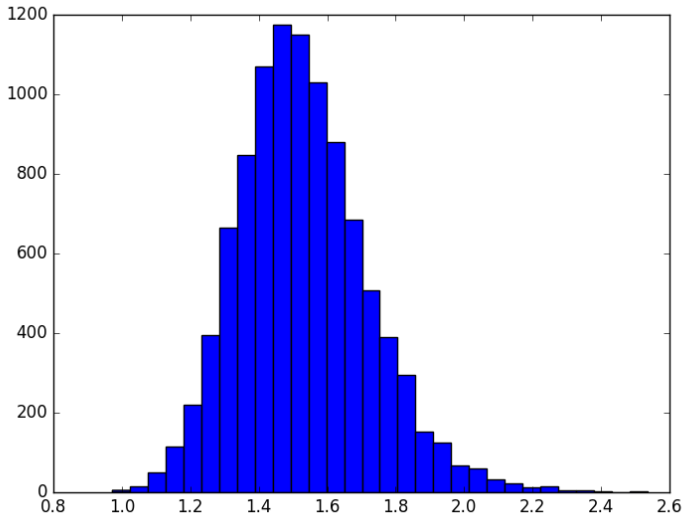


Figure: Gibbs sampling of the slope m.

- Gibbs is only one option for sampling. MCMC with Metropolis-Hastings, or Hamiltonian Monte Carlo, would also be perfectly viable.

Question: is this the most probable slope?

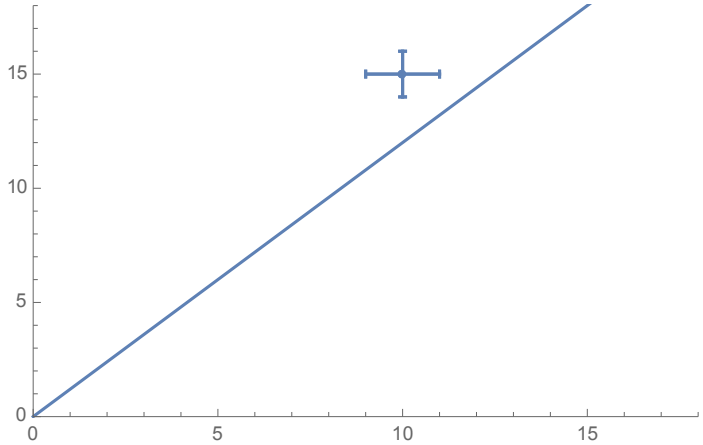


Figure: Noisy data

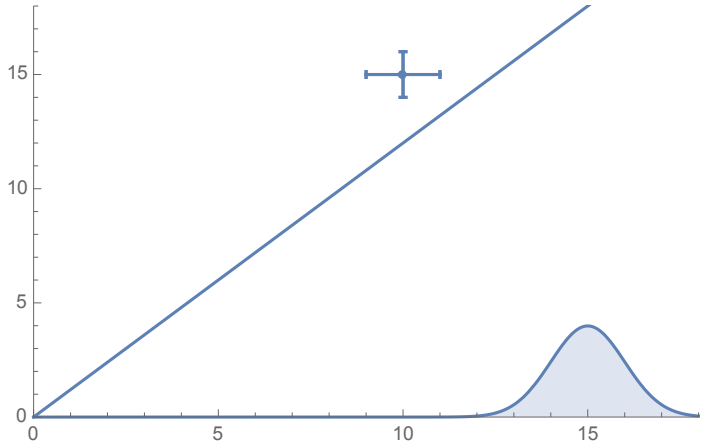


Figure: Yes! - there is a prior on $x \ldots$

Case Study. BPZ: Bayesian Photometric Redshifts

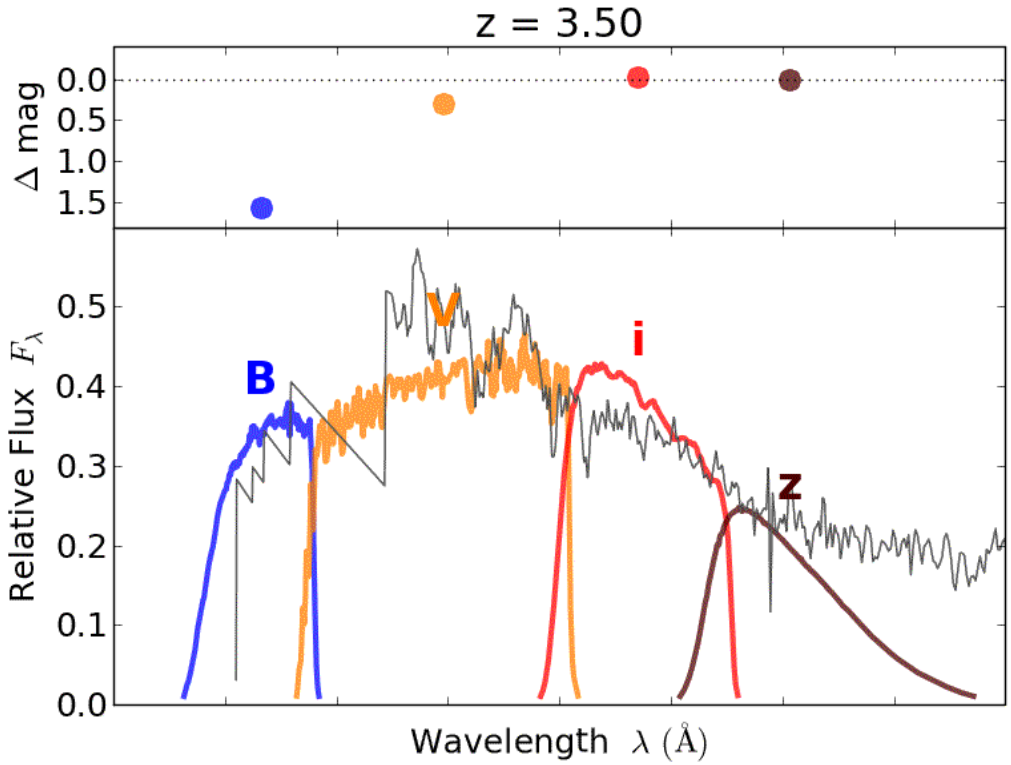


Figure: Spectrum and broad band fluxes

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

BPZ: Bayesian Photometric Redshifts

Goal

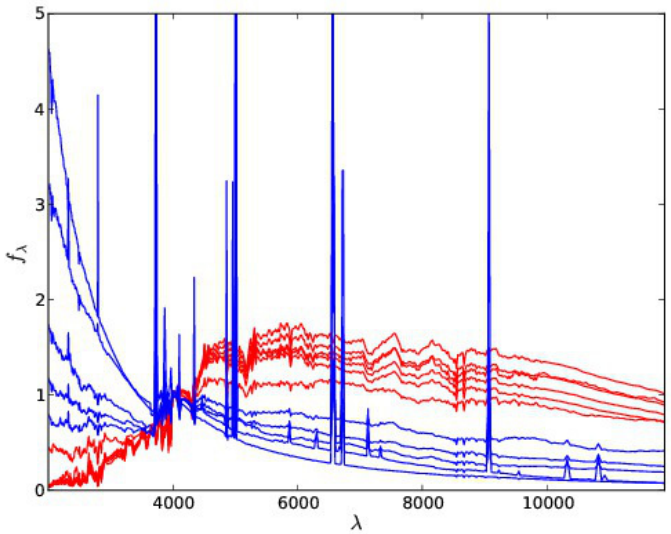
Obtain a posterior for the redshift of a galaxy given measurements of fluxes in some broadband filters (typically 5).

-

Model assumptions

Galaxy has a spectrum that is proportional to one of a set of template galaxies, but shifted in wavelength because of cosmological redshift.

-



Specify the model

Data:

$\hat{\mathbf{f}}$: vector of flux *measurements* $\hat{f}_\alpha$ in bands $\alpha = 1, \ldots N$

Parameters:

z: redshift

Latent variables:

$\mathbf{f}$: *True* f_α

T: template

a: amplitude of template 'brightness'

Posterior

- First write down what we want:
- $p(z|\hat{\mathbf{f}})$
- $= \sum_T p(z, T|\hat{\mathbf{f}})$ Marginalise over templates (discrete set)
- $= \sum_T \int da\, d\mathbf{f}\, p(z, T, a, \mathbf{f}|\hat{\mathbf{f}})$ and brightness and true fluxes
- $\propto \sum_T \int da\, d\mathbf{f}\, p(\hat{\mathbf{f}}|z, T, a, \mathbf{f})\, p(z, T, a, \mathbf{f})$
- $\propto \sum_T \int da\, d\mathbf{f}\, p(\hat{\mathbf{f}}|\mathbf{f})\, p(z, T, a, \mathbf{f})$. Measurement error only
- $\propto \sum_T \int da\, d\mathbf{f}\, p(\hat{\mathbf{f}}|\mathbf{f})\, p(\mathbf{f}|z, T, a) p(z, T, a)$
- $p(\mathbf{f}|z, T, a) = \delta(\mathbf{f} - a\,\mathbf{t}(T, z))$ where $\mathbf{t}(T, z)$ represents the fluxes of template T when shifted by z
- $p(z|\hat{\mathbf{f}}) \propto \sum_T \int da\, p(\hat{\mathbf{f}}|a\mathbf{t}(T, z))\, p(z|T, a)\, p(T, a)$
- $p(z|\hat{\mathbf{f}}) \propto \sum_T \int da\, p(\hat{\mathbf{f}}|a\mathbf{t}(T, z))\, p(z|T, a)\, p(T|a)\, p(a)$

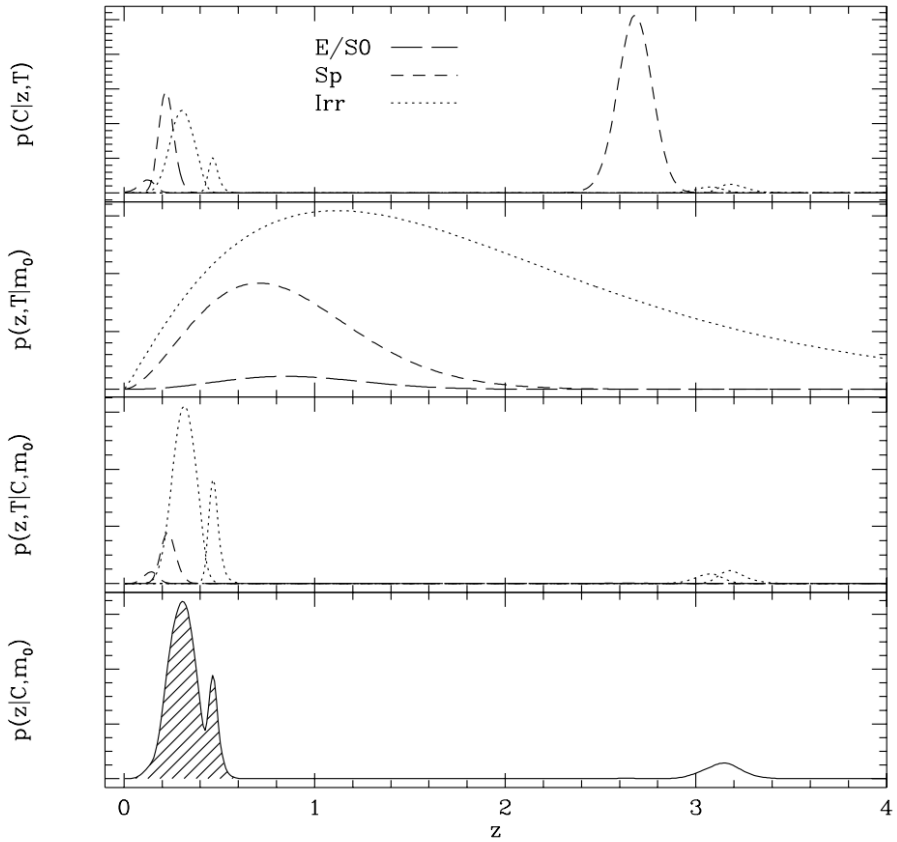
Posterior redshift distribution

- $p(z|\hat{\mathbf{f}}) \propto \sum_T \int da\, p(\hat{\mathbf{f}}|a\mathbf{t}(T,z))\, p(z|T,a)\, p(T|a)\, p(a)$
- **This identifies what we need to know**:
- $\mathbf{t}(T,z)$): Template T broadband fluxes (without renormalising), when redshifted by z
- $p(\hat{\mathbf{f}}|\mathbf{f})$: The error distribution for the fluxes (e.g. $\mathbf{f} \sim \mathcal{N}(\hat{\mathbf{f}}, \sigma_\alpha^2)$)
- $p(z|T,a)$: Redshift distribution of sources with template T and brightness a
- $p(T|a)$: Fraction of galaxies with template T, given a brightness a
- $p(a)$: Brightness distribution

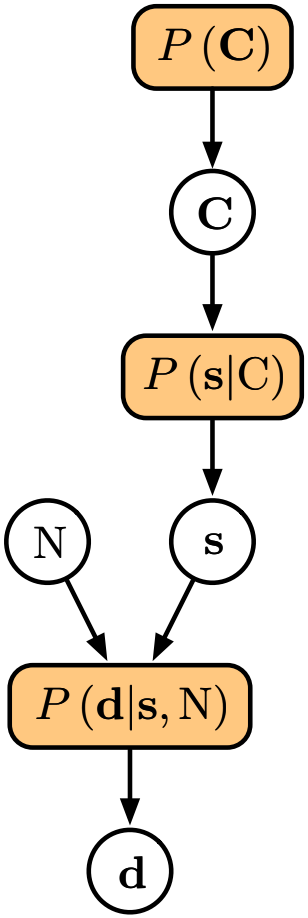
Frequentist vs. Bayesian

- $p(z|\hat{\mathbf{f}}) \propto \sum_T \int da\, p(\hat{\mathbf{f}}|a\mathbf{t}(T,z))\, p(z|T,a)\, p(T|a)\, p(a)$
- Frequentist interpretation of likelihood often differs from the Bayesian posterior only by the prior, which, if uniform, gives the same result:
- $p(\theta|D) \propto p(D|\theta)\, p(\theta)$
- **Here it is not so simple**. The maximum of the posterior (MAP: *maximum a posteriori*) is *not* the maximum likelihood, because T is marginalised over - the answer is very different

BPZ: likelihood and posterior



Weak Lensing BHM: Forward Model or Generative Model



C = Power Spectrum

s = shear map

N = noise variance in each pixel

d = noisy shear estimates in each pixel

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

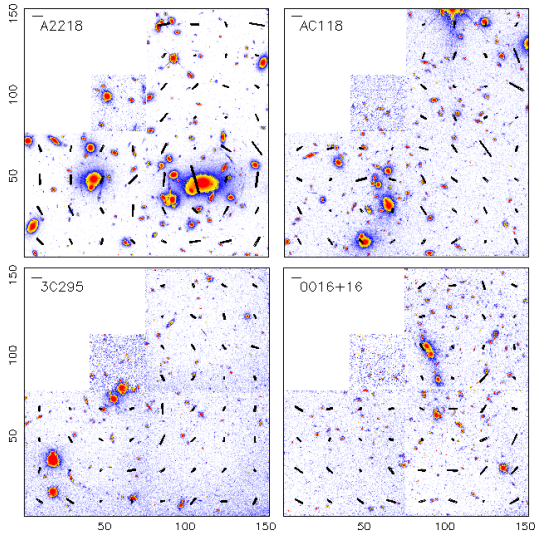


Figure: From Smail et al. 1997.

Latent parameters

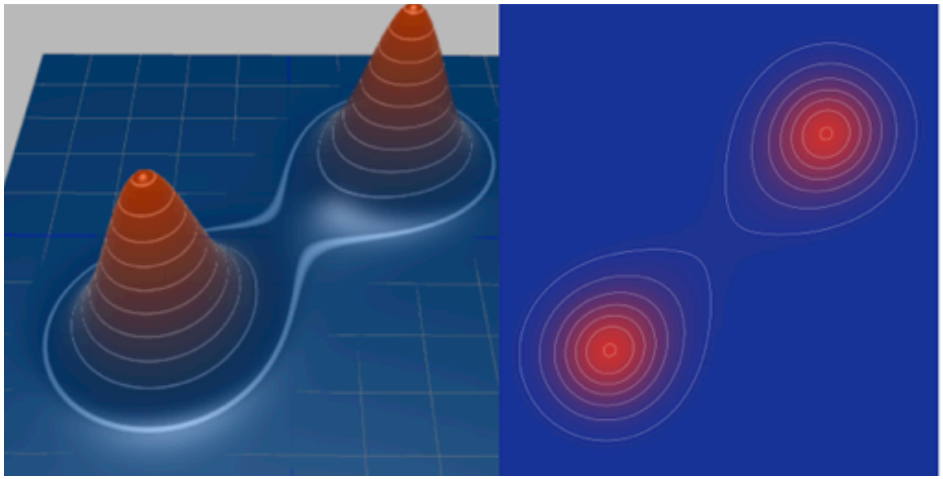
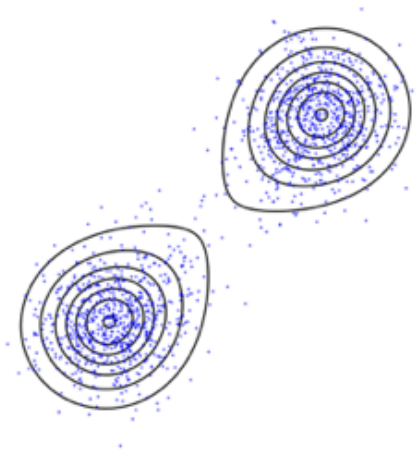
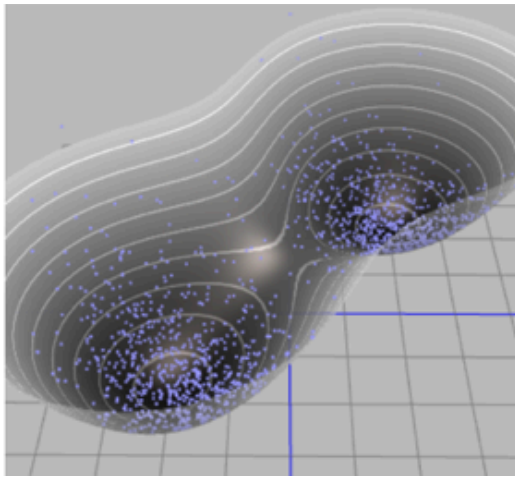
Each pixel in the map is a parameter

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

One million-dimensional probability distribution to calculate...

Sampling in *very* high dimensions

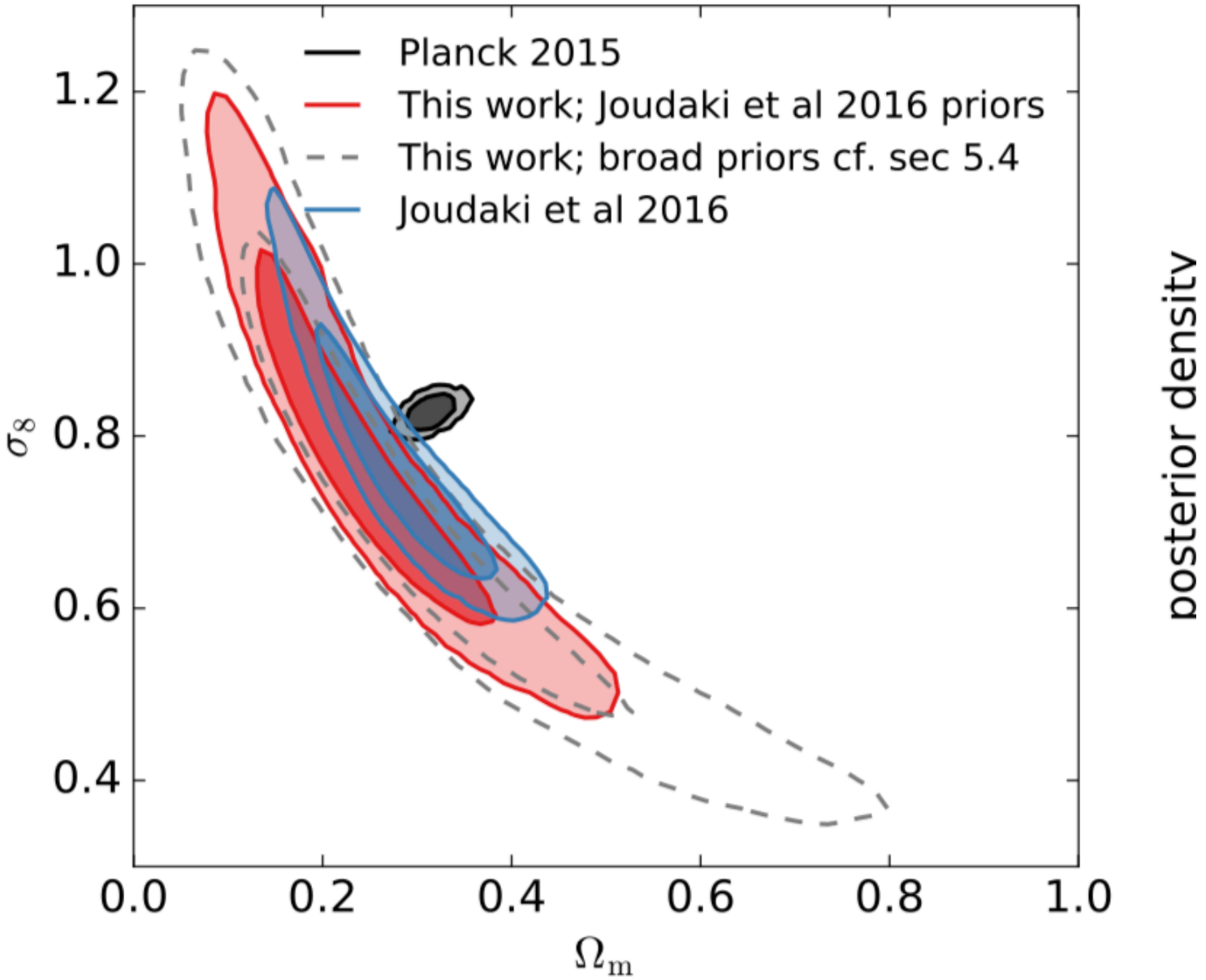
- MCMC: Metropolis-Hastings fails since it is very hard to devise an efficient proposal distribution
- Gibbs sampling: effective if conditional distributions are known
- Hamiltonian Monte Carlo (HMC) works in very high dimensions (e.g. using Stan)

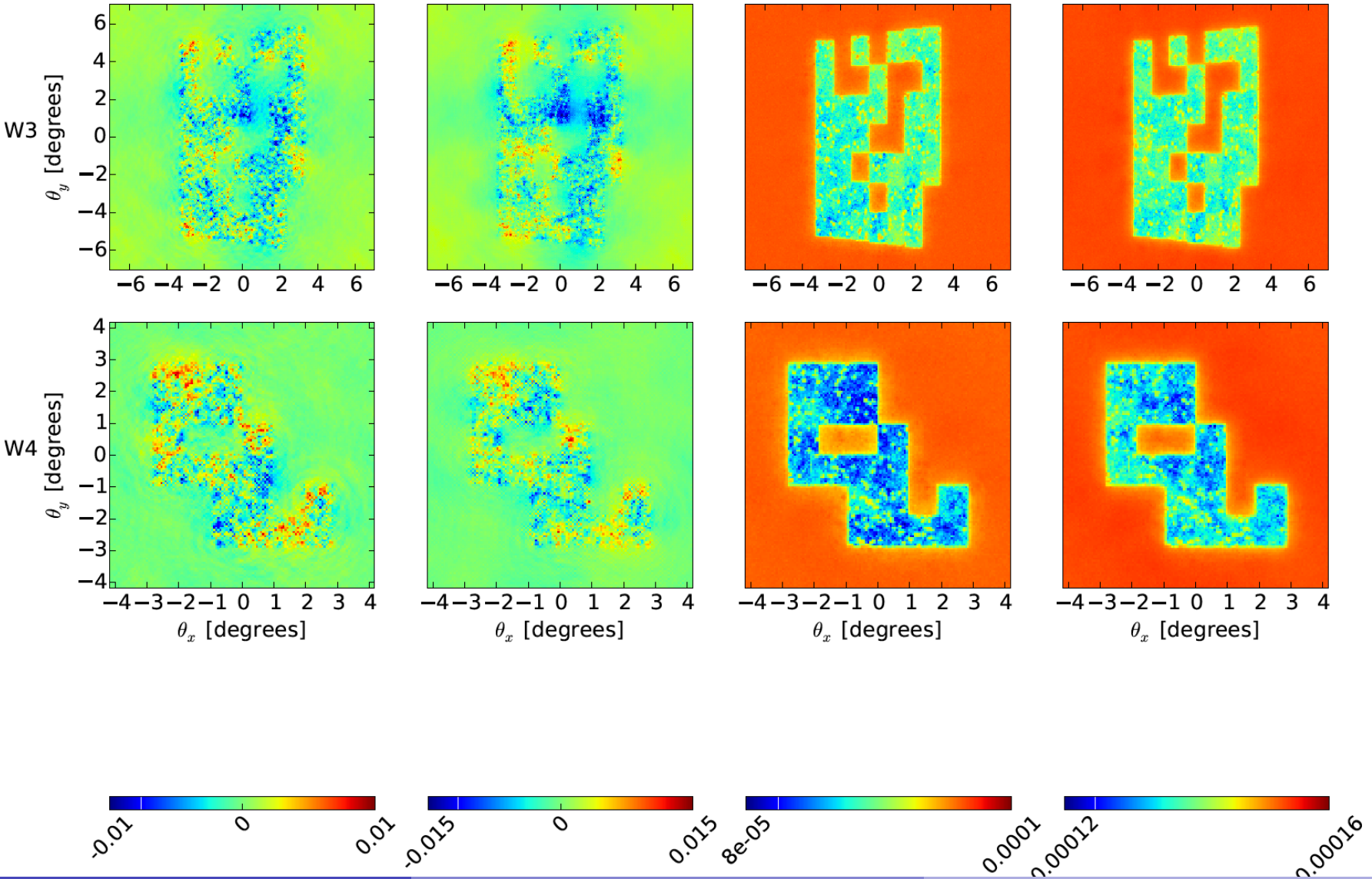
CFHTLenS

Alsing, AFH et al (2016). $\sim 250,000$ parameters; Gibbs sampling

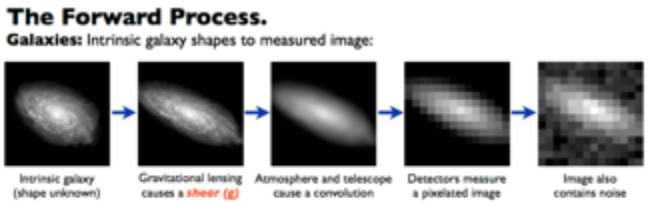
CFHTLenS results

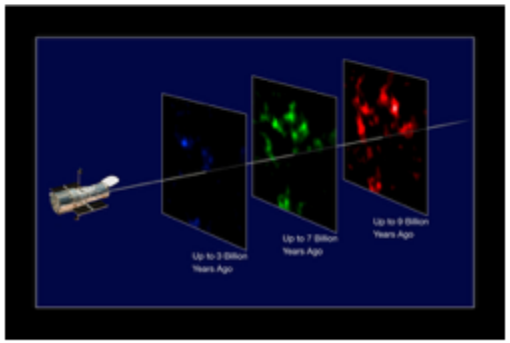


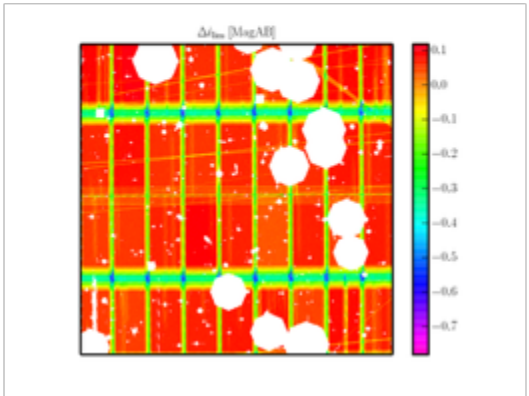
CFHTLenS matter maps



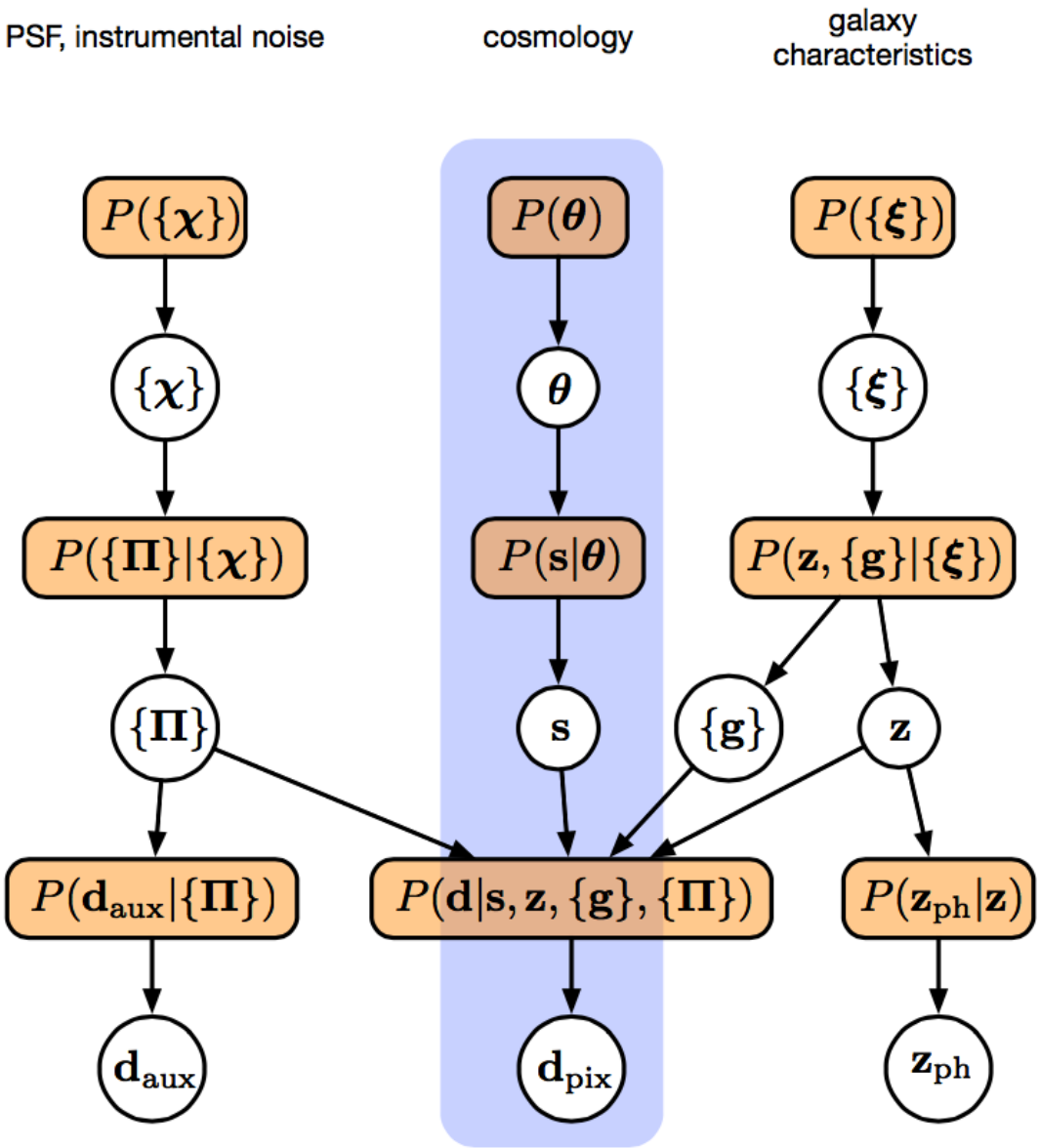
More Complications







Massive BHM


PSF, instrumental noise
cosmology
galaxy characteristics
$P(\{\chi\})$
$P(\theta)$
$P(\{\xi\})$
$\{\chi\}$
θ
$\{\xi\}$
$P(\{\Pi\}|\{\chi\})$
$P(\mathbf{s}|\theta)$
$P(\mathbf{z},\{\mathbf{g}\}|\{\xi\})$
$\{\Pi\}$
$\mathbf{s}$
$\{\mathbf{g}\}$
$\mathbf{z}$
$P(\mathbf{d}_{\rm aux}|\{\Pi\})$
$P(\mathbf{d}|\mathbf{s},\mathbf{z},\{\mathbf{g}\},\{\Pi\})$
$P(\mathbf{z}_{\rm ph}|\mathbf{z})$
$\mathbf{d}_{\rm aux}$
$\mathbf{d}_{\rm pix}$
$\mathbf{z}_{\rm ph}$

BORG and SDSS

Courtesy F. Leclercq.

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