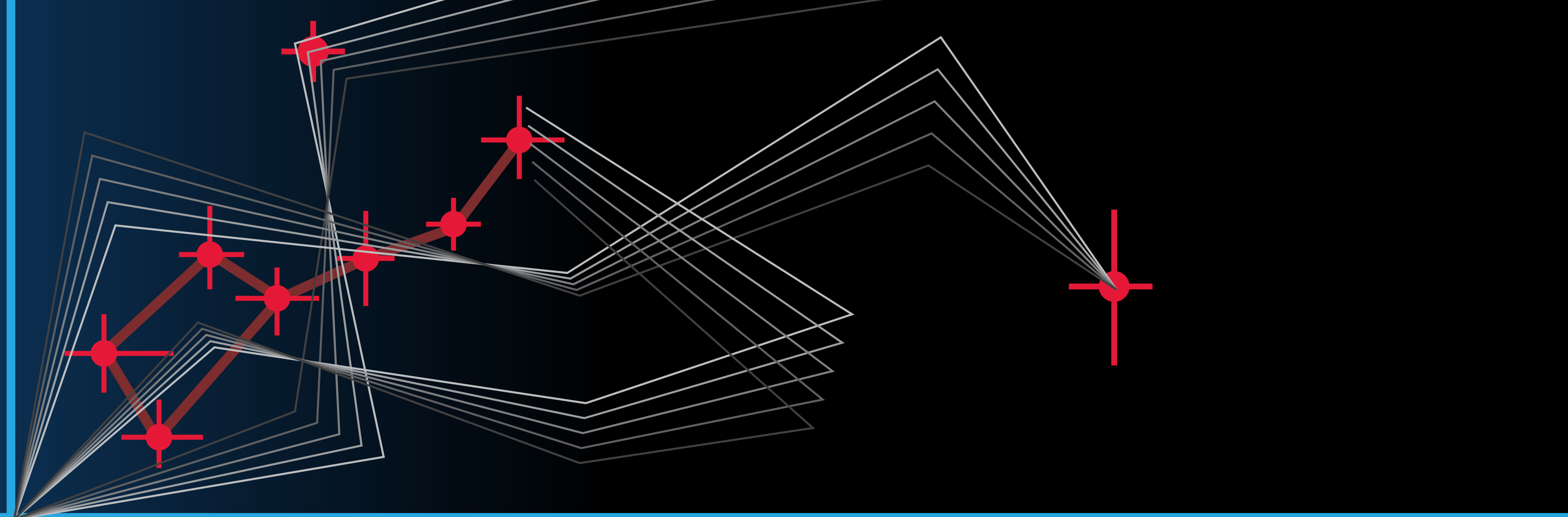
@R_Trotta

Bayesian inference: Principles and applications

Roberto Trotta - www.robertotrotta.com



ICIC
Imperial Centre
for Inference & Cosmology

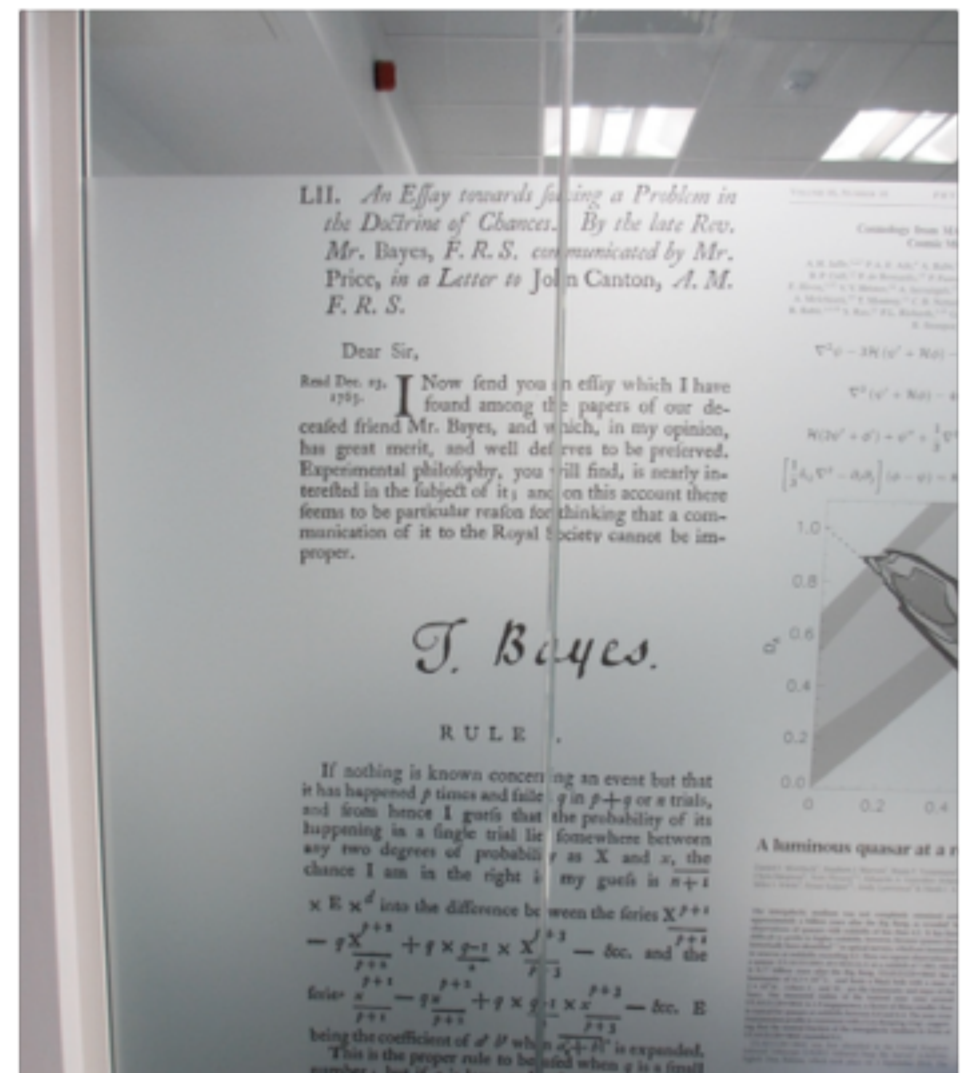
XII School of Cosmology
Cargese, Sept 16th 2014

Imperial College
London

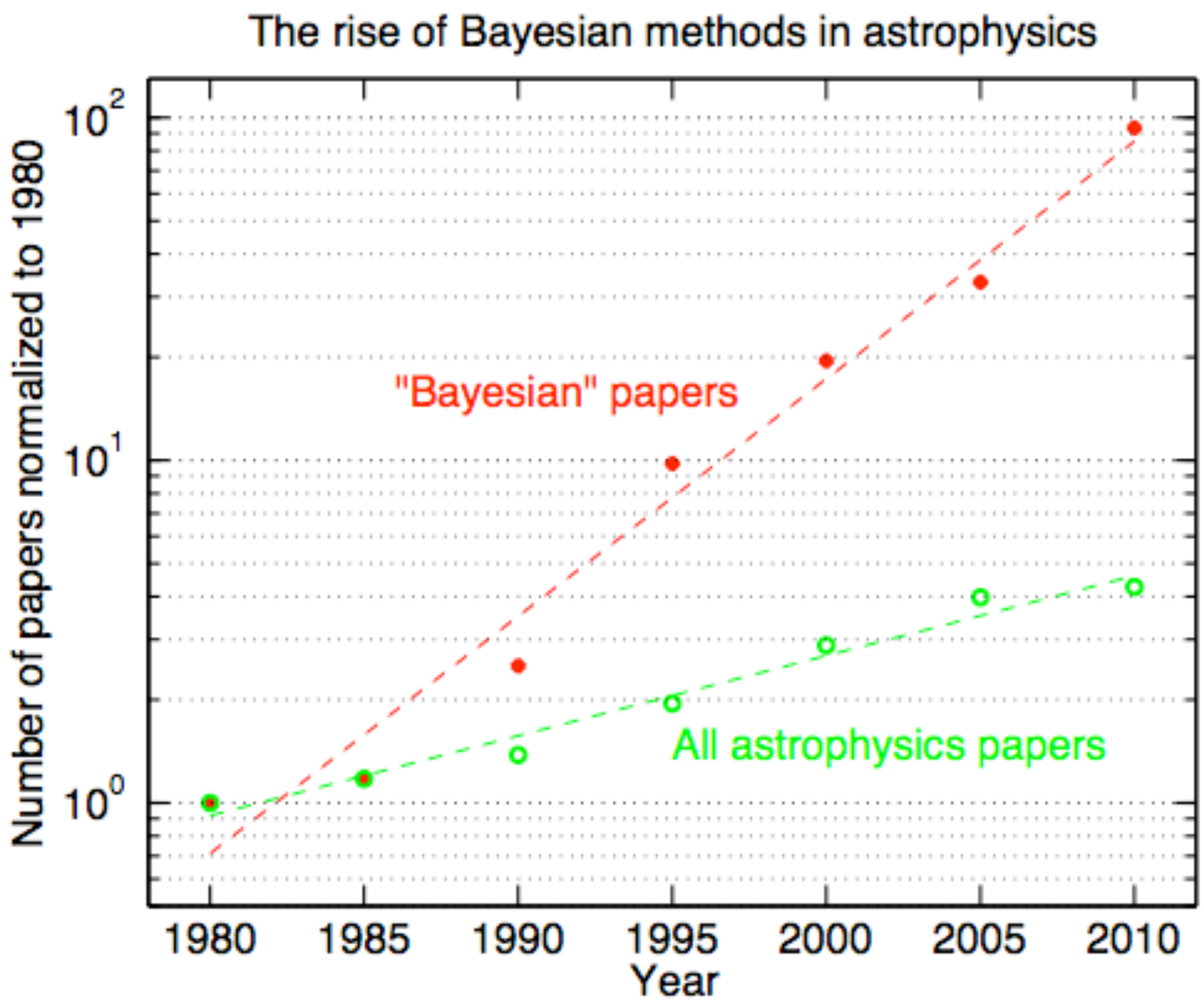
ICIC

Imperial Centre
for Inference & Cosmology

astro.ic.ac.uk/icic



Bayesian methods on the rise



Bayes Theorem

- Bayes' Theorem follows from the basic laws of probability: For two propositions A, B (not necessarily random variables!)

$$P(A|B)\ P(B) = P(A,B) = P(B|A)P(B)$$

$$P(A|B) = P(B|A)P(B)\ /\ P(A)$$

- Bayes' Theorem is simply **a rule to invert the order of conditioning of propositions.** This has PROFOUND consequences!

Bayes' theorem   P(A|B) = P(B|A)P(A) / P(B)

Imperial College London

posterior likelihood prior

$$P(\theta|d, I) = \frac{P(d|\theta, I)P(\theta|I)}{P(d|I)}$$

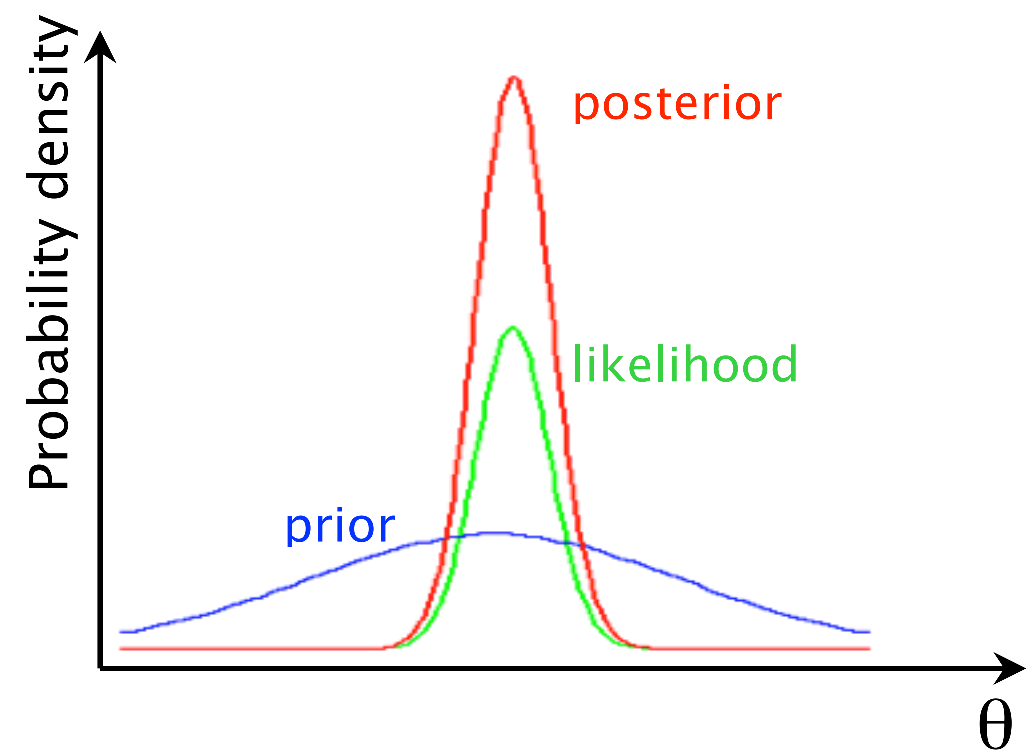
evidence

A → **θ:** parameters
B → **d:** data
I: any other external information, or the assumed model

For parameter inference it is sufficient to consider

$$P(\theta|d, I) \propto P(d|\theta, I)P(\theta|I)$$

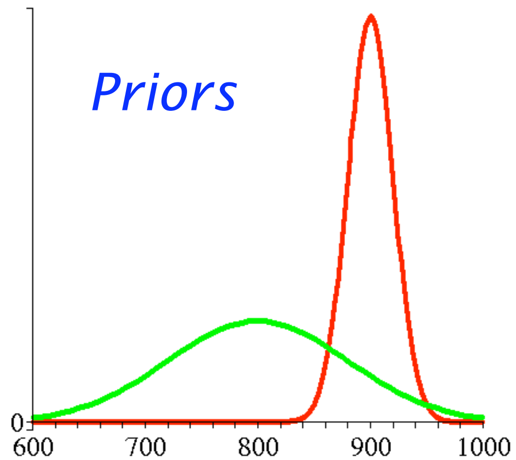
$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$



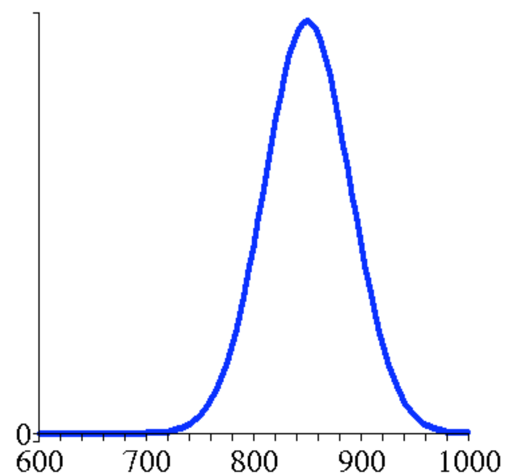
ICIC

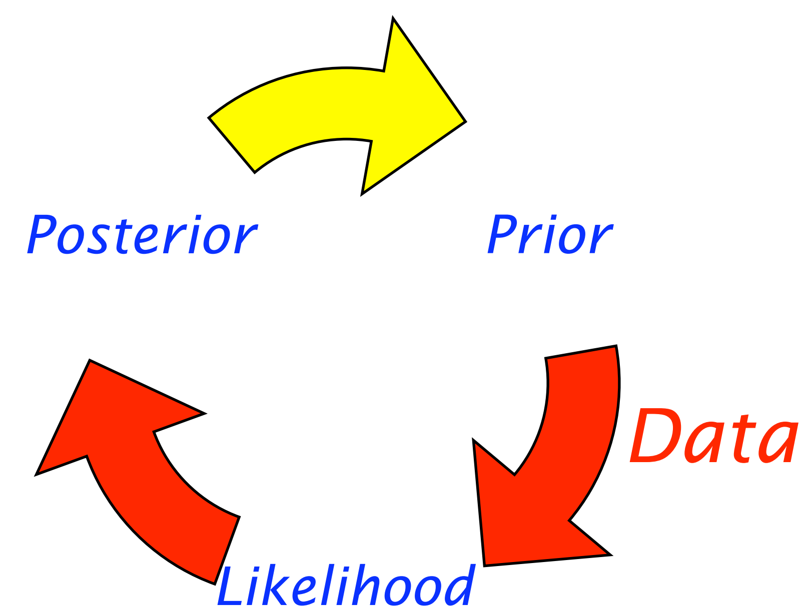
The matter with priors

- In parameter inference, prior dependence will **in principle** vanish for strongly constraining data.
  A sensitivity analysis is mandatory for all Bayesian methods!

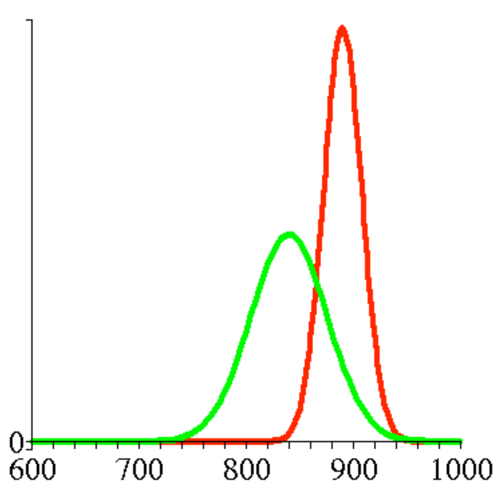


Likelihood (1 datum)

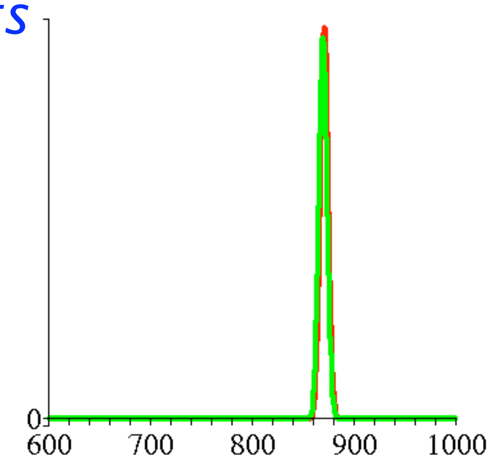




Posterior after 1 datum



Posterior after 100 data points



Inference in many dimensions

Imperial College London

Usually our parameter space is multi-dimensional: how should we report inferences for one parameter at the time?

BAYESIAN

Marginal posterior:

$$P(\theta_1|D) = \int L(\theta_1, \theta_2) p(\theta_1, \theta_2) d\theta_2$$

FREQUENTIST

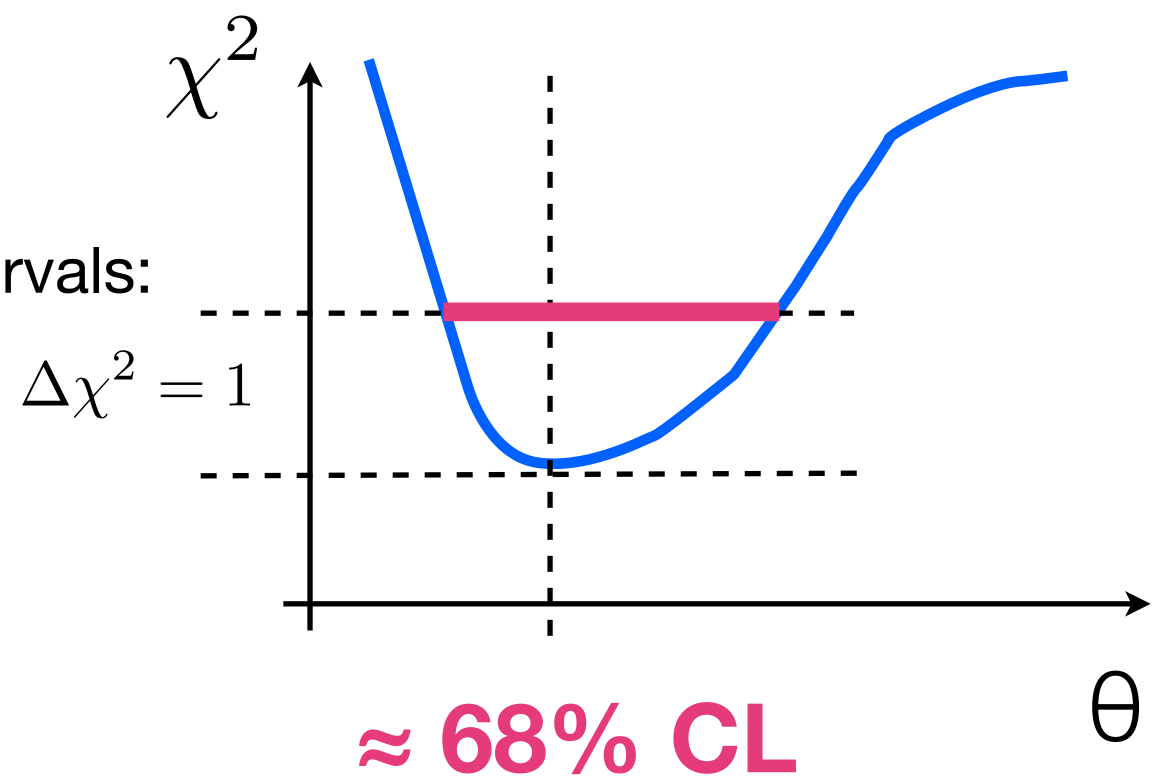
Profile likelihood:

$$L(\theta_1) = max_{\theta_2} L(\theta_1, \theta_2)$$

Confidence intervals: Frequentist approach

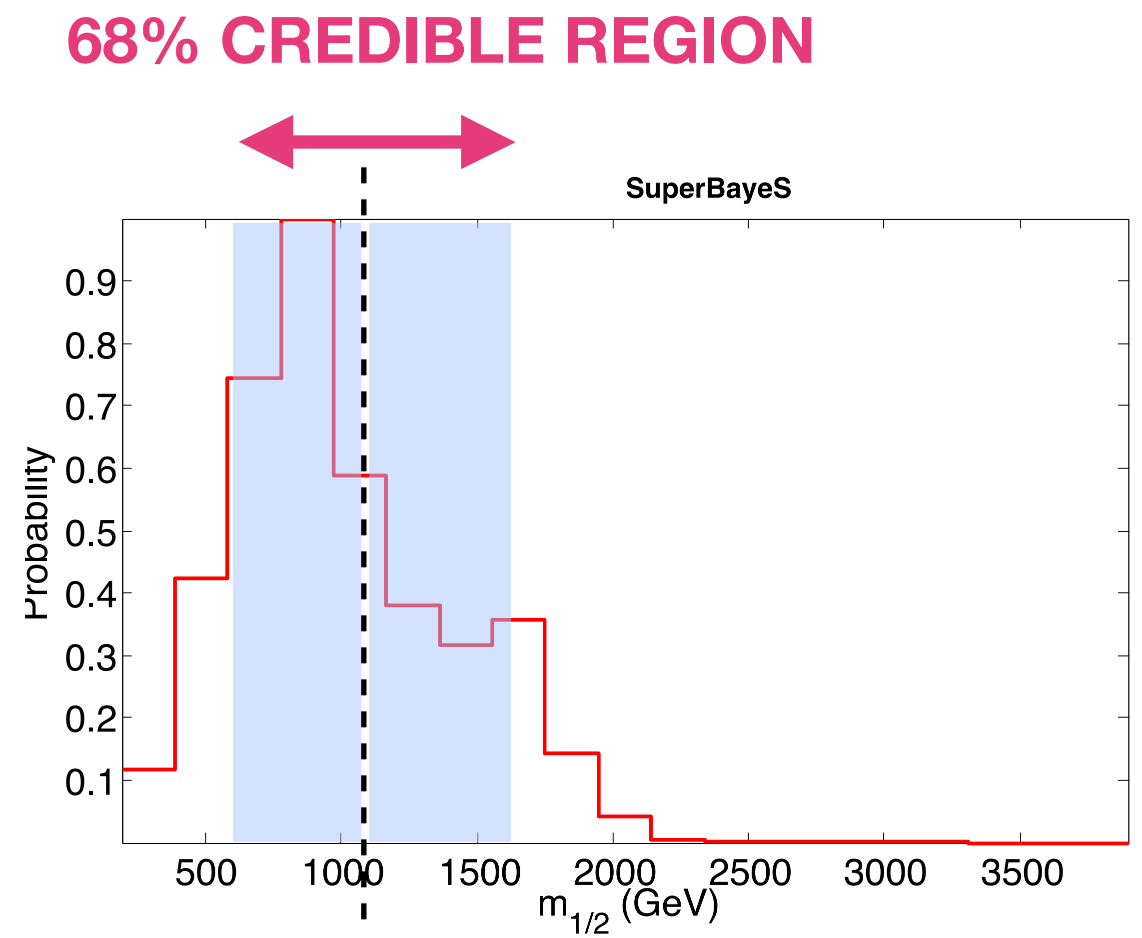
Imperial College London

- **Likelihood-based methods:** determine the best fit parameters by finding the minimum of -2Log(Likelihood) = chi-squared
  - Analytical for Gaussian likelihoods
  - Generally numerical
  - Steepest descent, MCMC, ...
- Determine approximate confidence intervals: Local Δ(chi-squared) method



Credible regions: Bayesian approach

- Use the prior to define a metric on parameter space.
- **Bayesian methods:** the best-fit has no special status. Focus on region of large posterior probability mass instead.
    - Markov Chain Monte Carlo (MCMC)
    - Nested sampling
    - Hamiltonian MC
- Determine posterior credible regions: e.g. symmetric interval around the mean containing 68% of samples



Marginalization vs Profiling

- Marginalisation of the posterior pdf (Bayesian) and profiling of the likelihood (frequentist) give exactly identical results for the linear Gaussian case.
- But: THIS IS NOT GENERICALLY TRUE!
- Sometimes, it might be useful and informative to look at both.

Marginalization vs profiling (maximising)

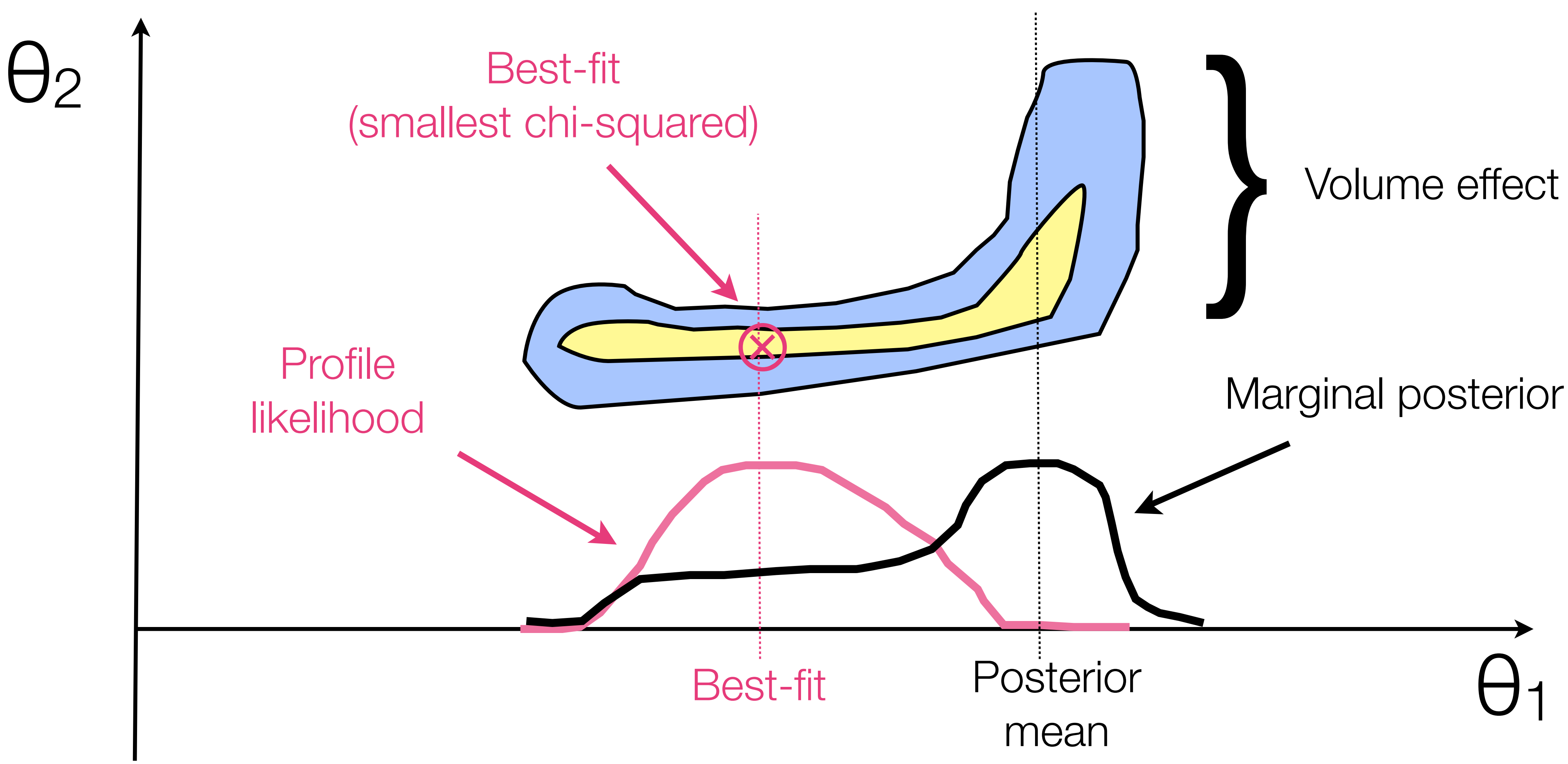
Imperial College London

Marginal posterior:

$$P(\theta_1|D) = \int L(\theta_1, \theta_2) p(\theta_1, \theta_2) d\theta_2$$

Profile likelihood:

$$L(\theta_1) = max_{\theta_2} L(\theta_1, \theta_2)$$



(2D plot depicts likelihood contours - prior assumed flat over wide range)

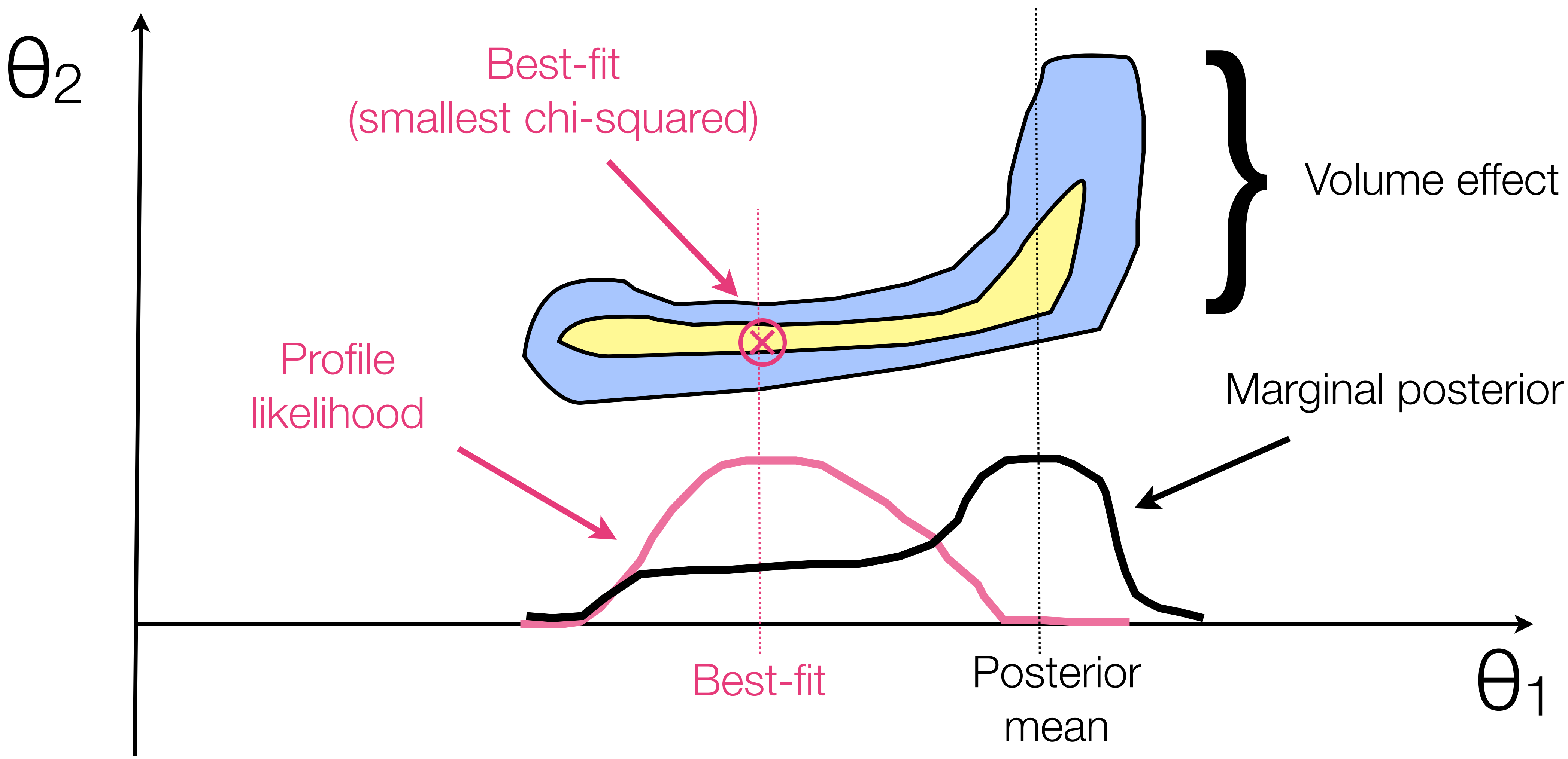
Marginalization vs profiling (maximising)

Physical analogy: (thanks to Tom Loredo)

Heat: $Q = \int c_V(x)T(x)dV$

Likelihood = hottest hypothesis
Posterior = hypothesis with most heat

Posterior: $P \propto \int p(\theta)L(\theta)d\theta$



(2D plot depicts likelihood contours - prior assumed flat over wide range)

What does x=1.00±0.01 mean?

Imperial College London

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

$$\text{Notation}: \quad x \sim N(\mu, \sigma^2)$$

- **Frequentist statistics (Fisher, Neymann, Pearson):**
  E.g., estimation of the mean μ of a Gaussian distribution from a list of observed samples x_1, x_2, x_3...
  The sample mean is the Maximum Likelihood estimator for μ:

  $\mu_{ML} = X_{av} = (x_1 + x_2 + x_3 + \ldots x_N)/N$
- **Key point:**
  in $P(X_{av})$, X_{av} is a random variable, i.e. one that takes on different values across an ensemble of infinite (imaginary) identical experiments. X_{av} is distributed according to $X_{av} \sim N(\mu, \sigma^2/N)$ **for a fixed true μ**
  The distribution applies to imaginary replications of data.

What does $x=1.00\pm0.01$ mean?

- **Frequentist statistics (Fisher, Neymann, Pearson):**
  The final result for the confidence interval for the mean

$$P(\mu_{ML} - \sigma/N^{1/2} < \mu < \mu_{ML} + \sigma/N^{1/2}) = 0.683$$

- This means:
  If we were to repeat this measurements many times, and obtain a 1-sigma distribution for the mean, the true value μ would lie inside the so-obtained intervals 68.3% of the time
- This is not the same as saying: “The probability of μ to lie within a given interval is 68.3%”. This statement only follows from using Bayes theorem.

What does x=1.00±0.01 mean?

Imperial College London

- **Bayesian statistics (Laplace, Gauss, Bayes, Bernouilli, Jaynes):**

  After applying Bayes therorem $P(\mu | X_{av})$ describes the distribution of our degree of belief about the value of μ given the information at hand, i.e. the observed data.

- Inference is conditional only on the observed values of the data.
- There is no concept of repetition of the experiment.

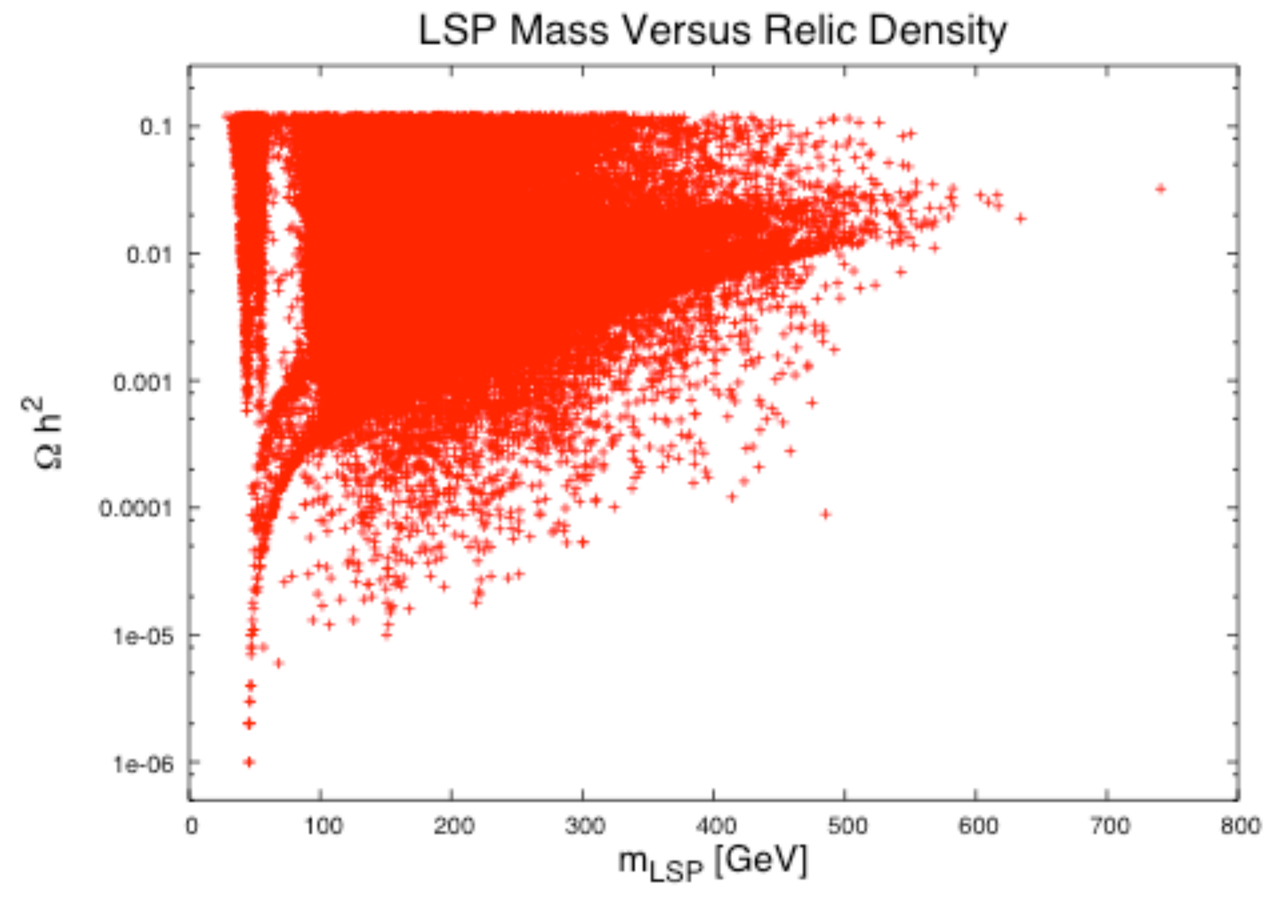
Markov Chain Monte Carlo

Exploration with “random scans”

Imperial College London

- Points accepted/rejected in a in/out fashion (e.g., 2-sigma cuts)
- No statistical measure attached to density of points: no probabilistic interpretation of results possible, although the temptation cannot be resisted...
- Inefficient in high dimensional parameters spaces (D>5)
- **HIDDEN PROBLEM:** Random scan explore only a very limited portion of the parameter space!

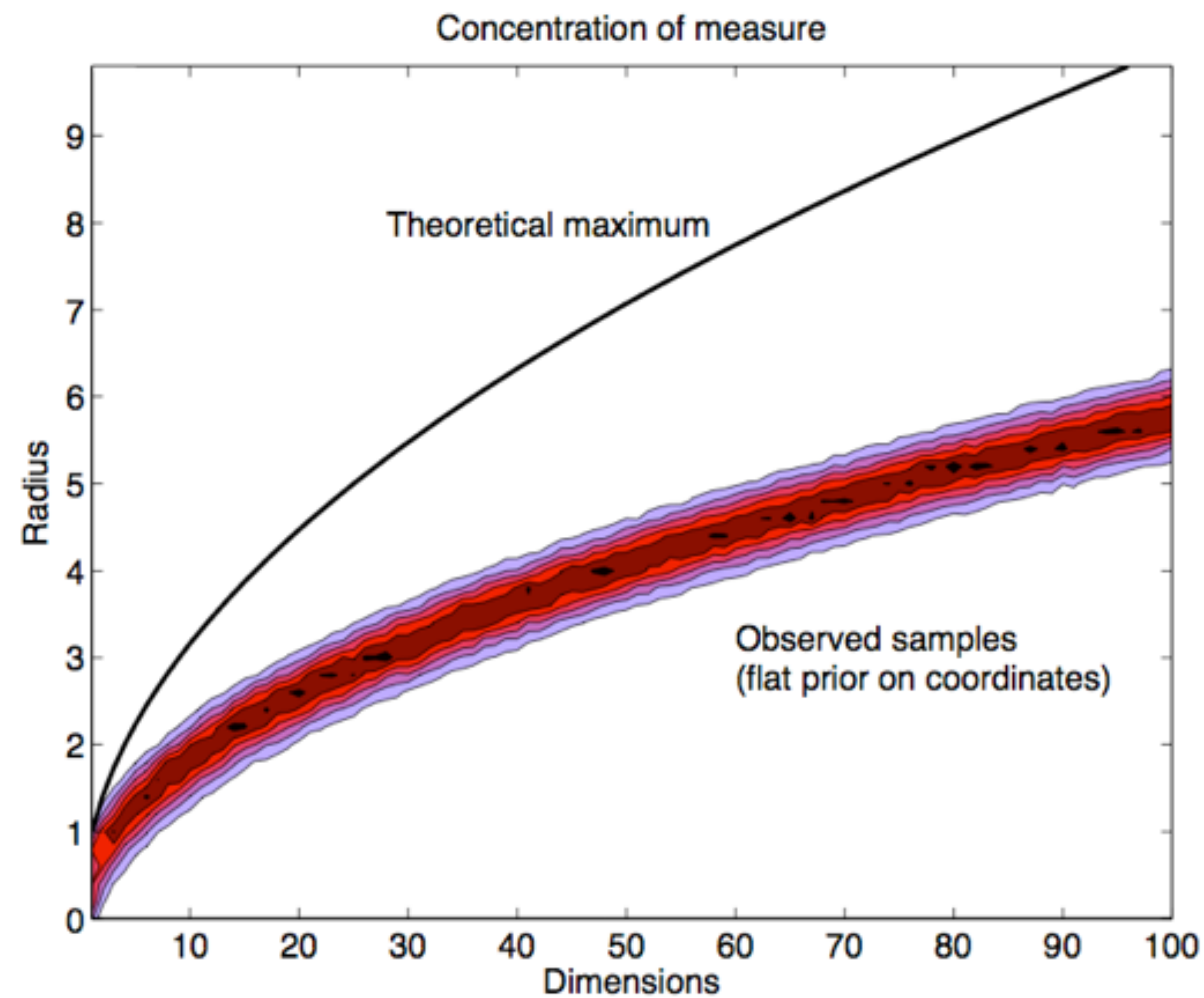
One recent example:
Berger et al (0812.0980)
pMSSM scans
(20 dimensions)

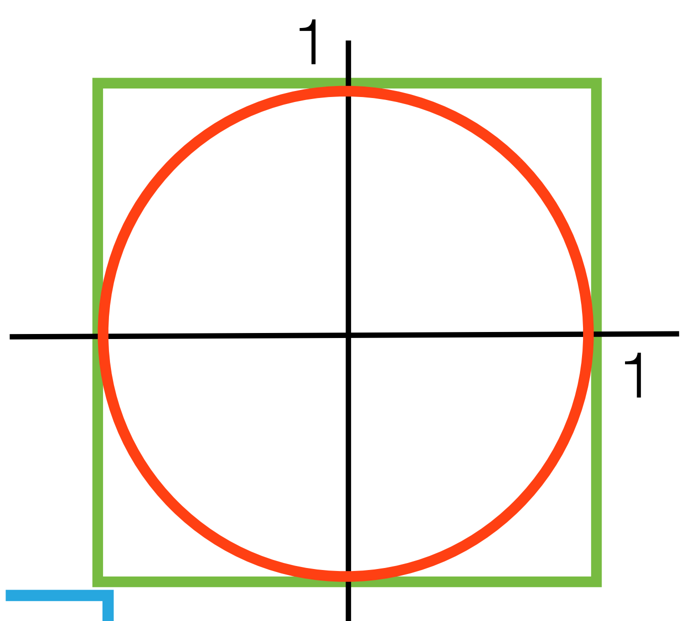


Random scans explore only a small fraction of the parameter space

Imperial College London

- "Random scans" of a high-dimensional parameter space only probe a very limited sub-volume: this is **the concentration of measure phenomenon.**
- **Statistical fact:** the norm of *D* draws from U[0,1] concentrates around $(D/3)^{1/2}$ with constant variance





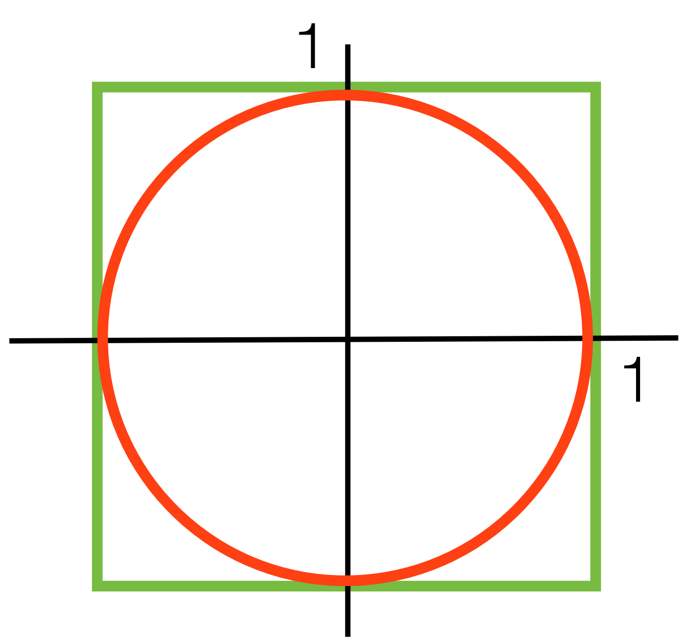
ICIC

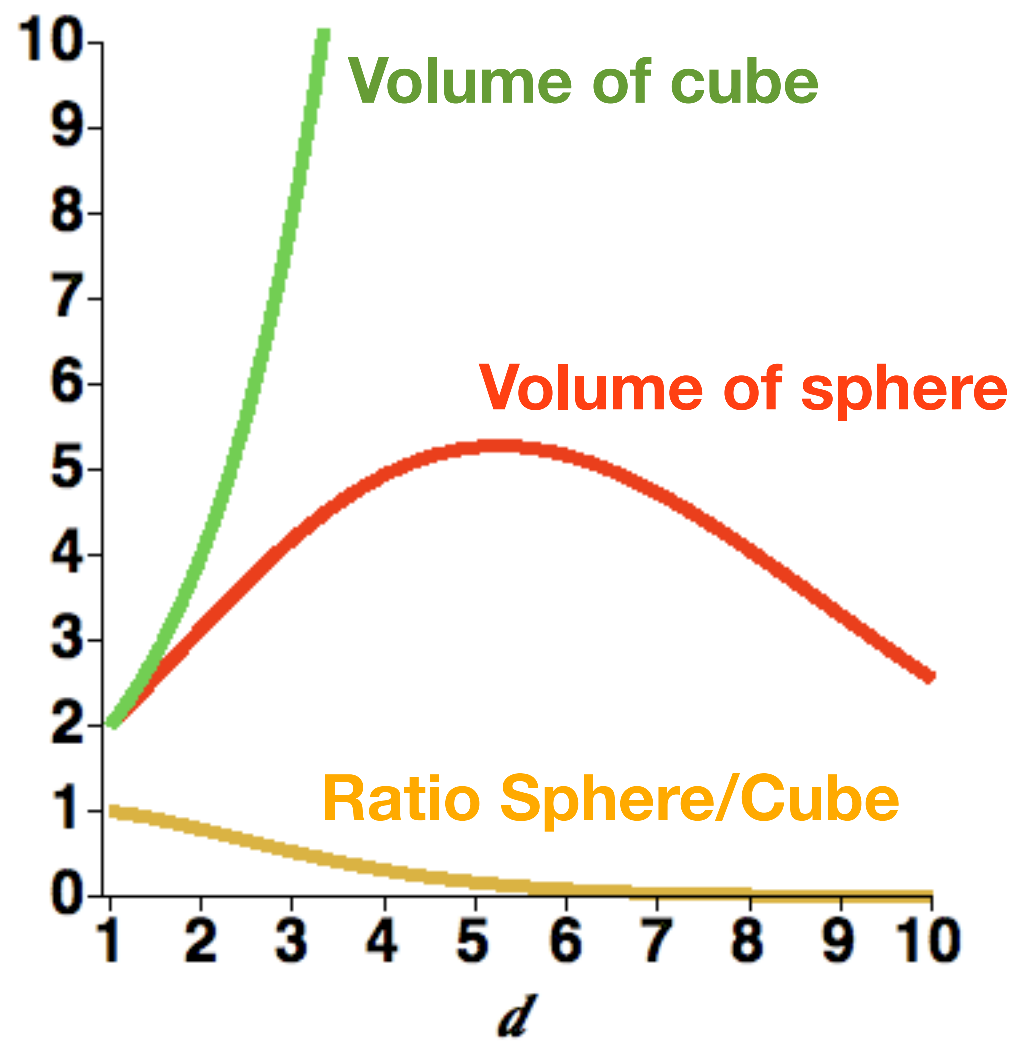
Geometry in high-D spaces

Imperial College London

- **Geometrical fact:** in *D* dimensions, most of the volume is near the boundary. The volume inside the spherical core of *D*-dimensional cube is negligible.

Together, these two facts mean that random scan only explore a very small fraction of the available parameter space in high-dimesional models.





Key advantages of the Bayesian approach

Imperial College London

- **Efficiency:** computational effort scales ~ N rather than k^N as in grid-scanning methods. Orders of magnitude improvement over grid-scanning.
- **Marginalisation:** integration over hidden dimensions comes for free.
- **Inclusion of nuisance parameters:** simply include them in the scan and marginalise over them.
- **Pdf's for derived quantities**: probabilities distributions can be derived for any function of the input variables

The general solution

Imperial College London

$$P(\theta|d, I) \propto P(d|\theta, I)P(\theta|I)$$

- Once the RHS is defined, how do we evaluate the LHS?
- Analytical solutions exist only for the simplest cases (e.g. Gaussian linear model)
- Cheap computing power means that numerical solutions are often just a few clicks away!
- Workhorse of Bayesian inference: Markov Chain Monte Carlo (MCMC) methods. A procedure to generate a list of samples from the posterior.

MCMC estimation

$$P(\theta|d,I) \propto P(d|\theta,I)P(\theta|I)$$

- A Markov Chain is a list of samples θ_1, θ_2, θ_3,... whose density reflects the (unnormalized) value of the posterior
- A MC is a sequence of random variables whose *(n+1)*-th elements only depends on the value of the *n*-th element
- **Crucial property:** a Markov Chain converges to a stationary distribution, i.e. one that does not change with time. In our case, the posterior.
- From the chain, expectation values wrt the posterior are obtained very simply:

$$\langle\theta\rangle = \int d\theta P(\theta|d)\theta \approx \frac{1}{N}\sum_i \theta_i$$

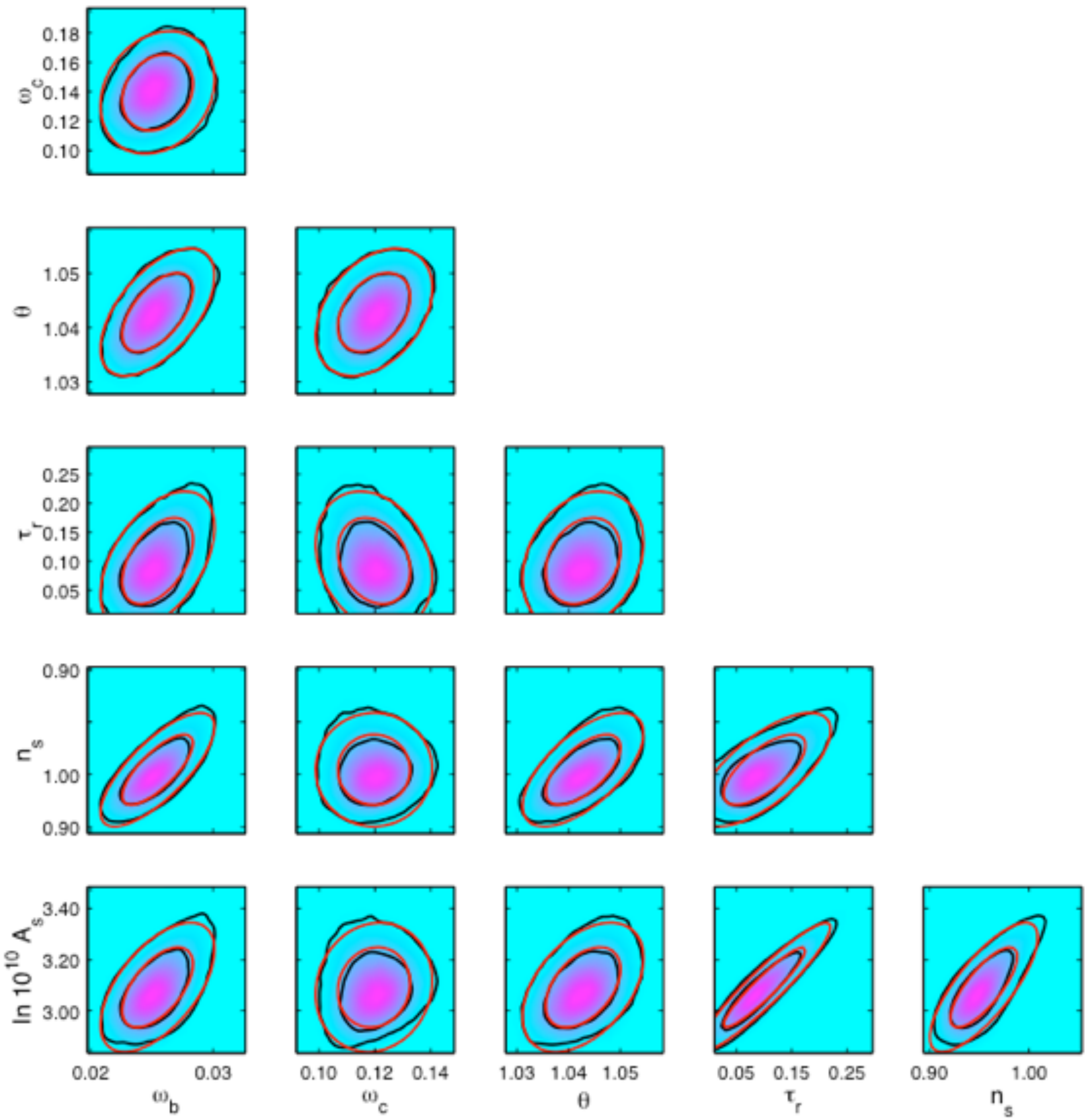
$$\langle f(\theta)\rangle = \int d\theta P(\theta|d) f(\theta) \approx \frac{1}{N}\sum_i f(\theta_i)$$

Reporting inferences

- **Once P(θ|d, I) found, we can report inference by:**
  - Summary statistics (best fit point, average, mode)
  - Credible regions (e.g. shortest interval containing 68% of the posterior probability for θ). Warning: this has **not** the same meaning as a frequentist confidence interval! (Although the 2 might be formally identical)
  - Plots of the marginalised distribution, integrating out nuisance parameters (i.e. parameters we are not interested in). This generalizes the propagation of errors:

$$P(\theta|d,I) = \int d\phi P(\theta,\phi|d,I)$$

Gaussian case

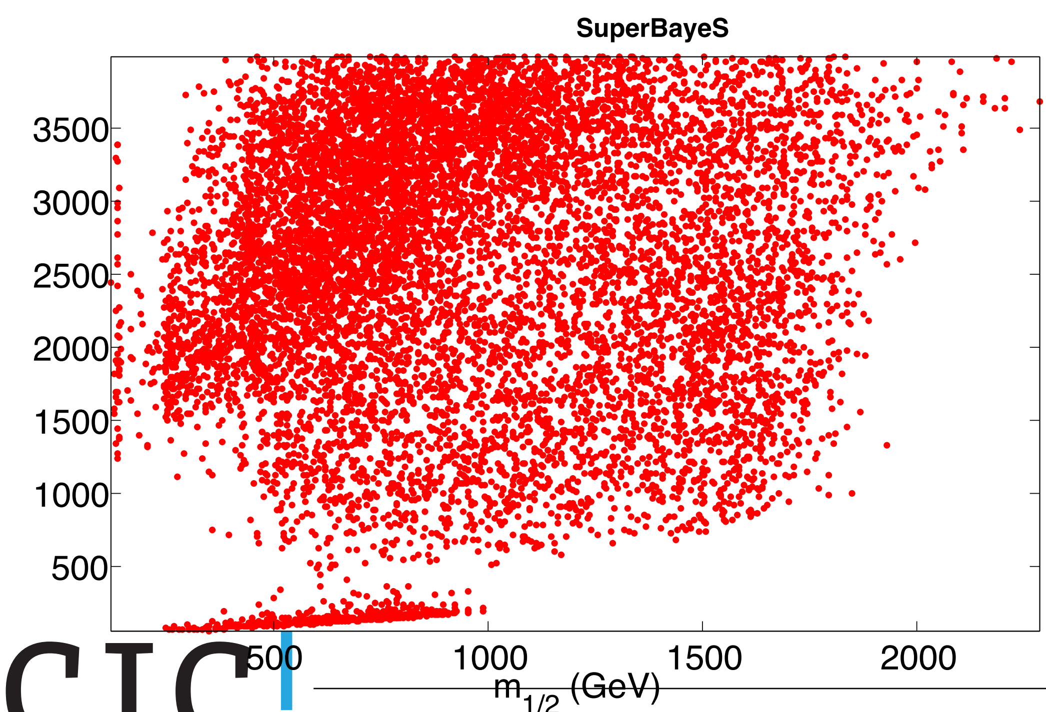


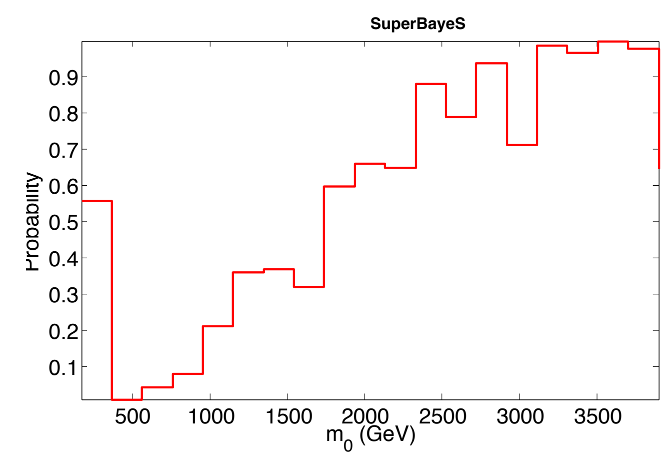
MCMC estimation

Imperial College London

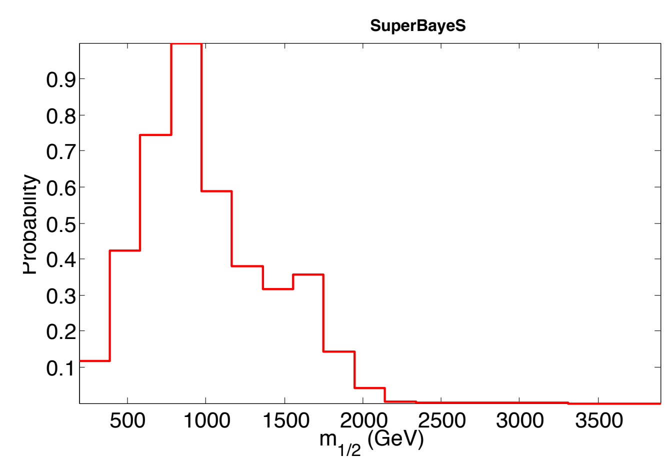
- **Marginalisation becomes trivial:** create bins along the dimension of interest and simply count samples falling within each bins ignoring all other coordinates
- Examples (from **superbayes.org**) :

2D distribution of samples from joint posterior





1D marginalised posterior (along y)



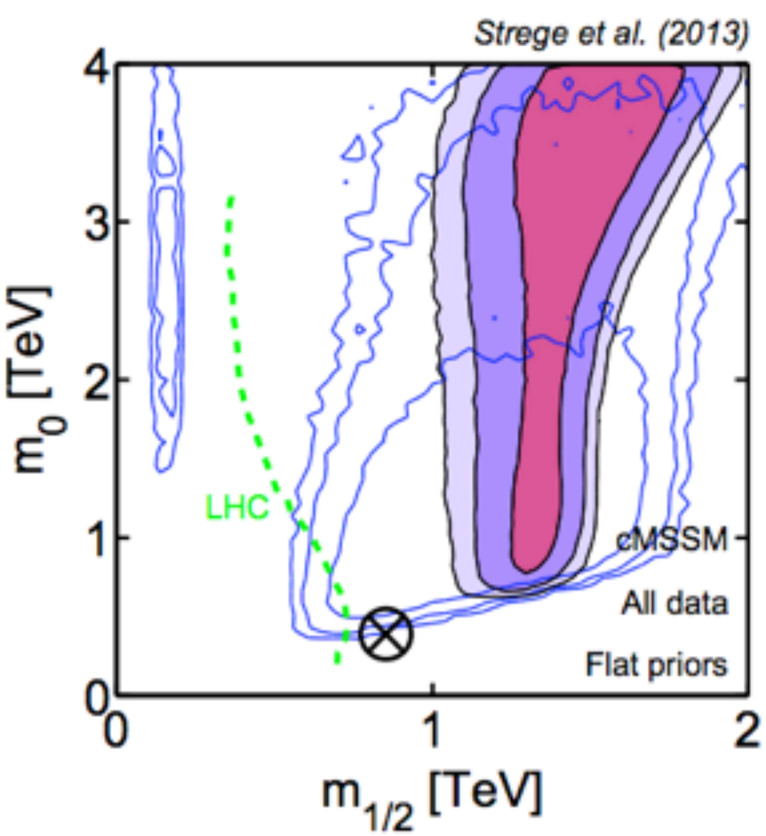
1D marginalised posterior (along x)

ICIC

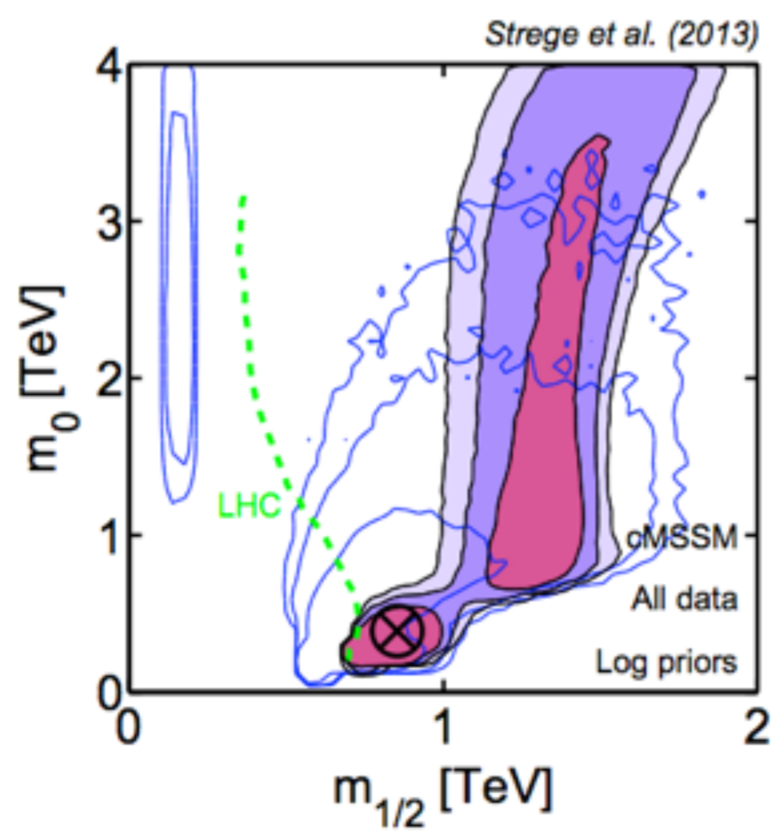
Non-Gaussian example

Imperial College London

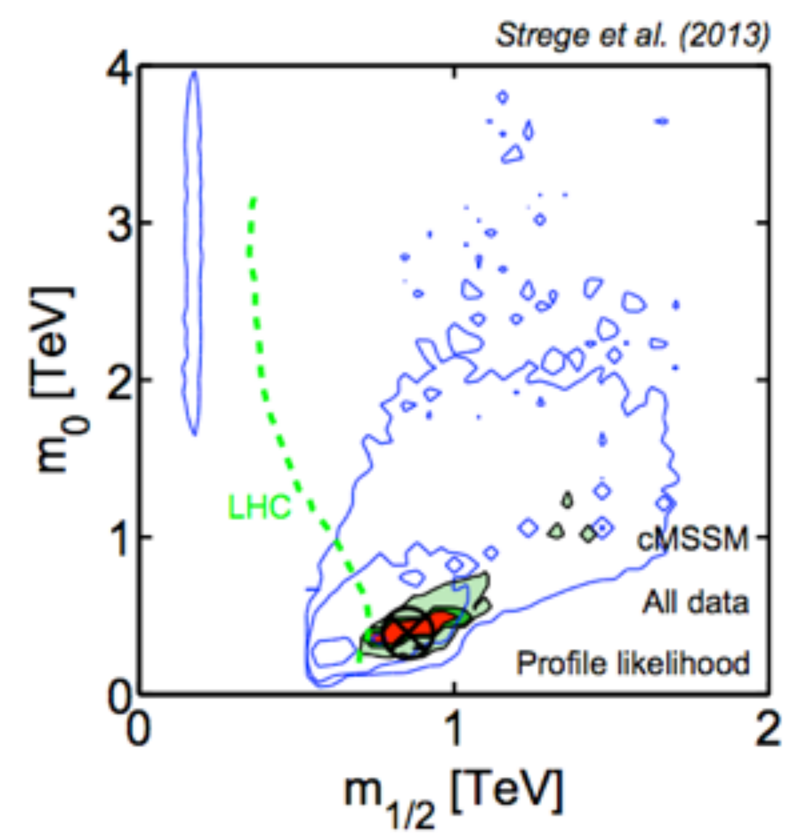
Bayesian posterior ("flat priors")



Bayesian posterior ("log priors")



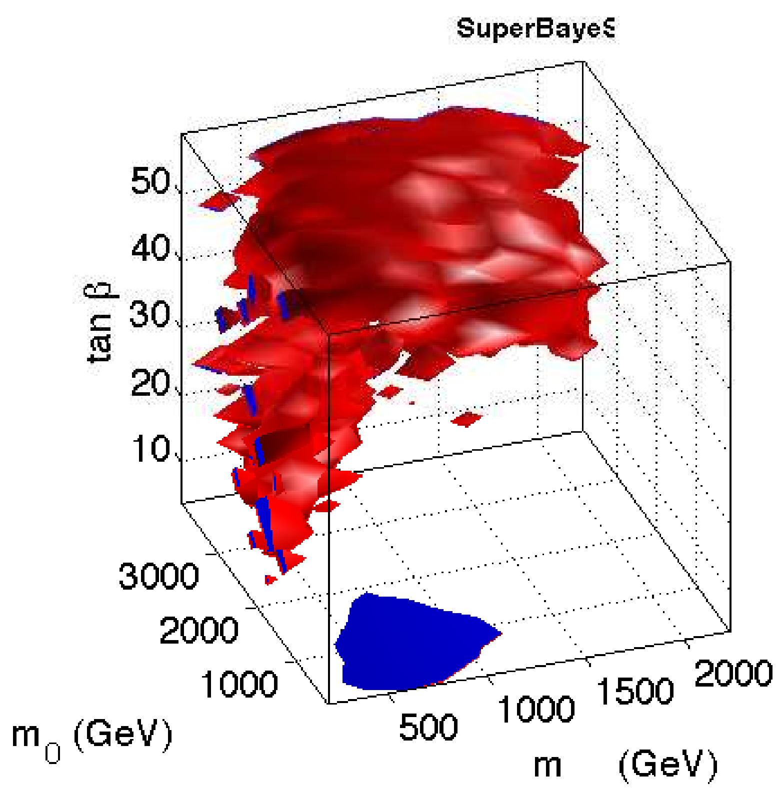
Profile likelihood

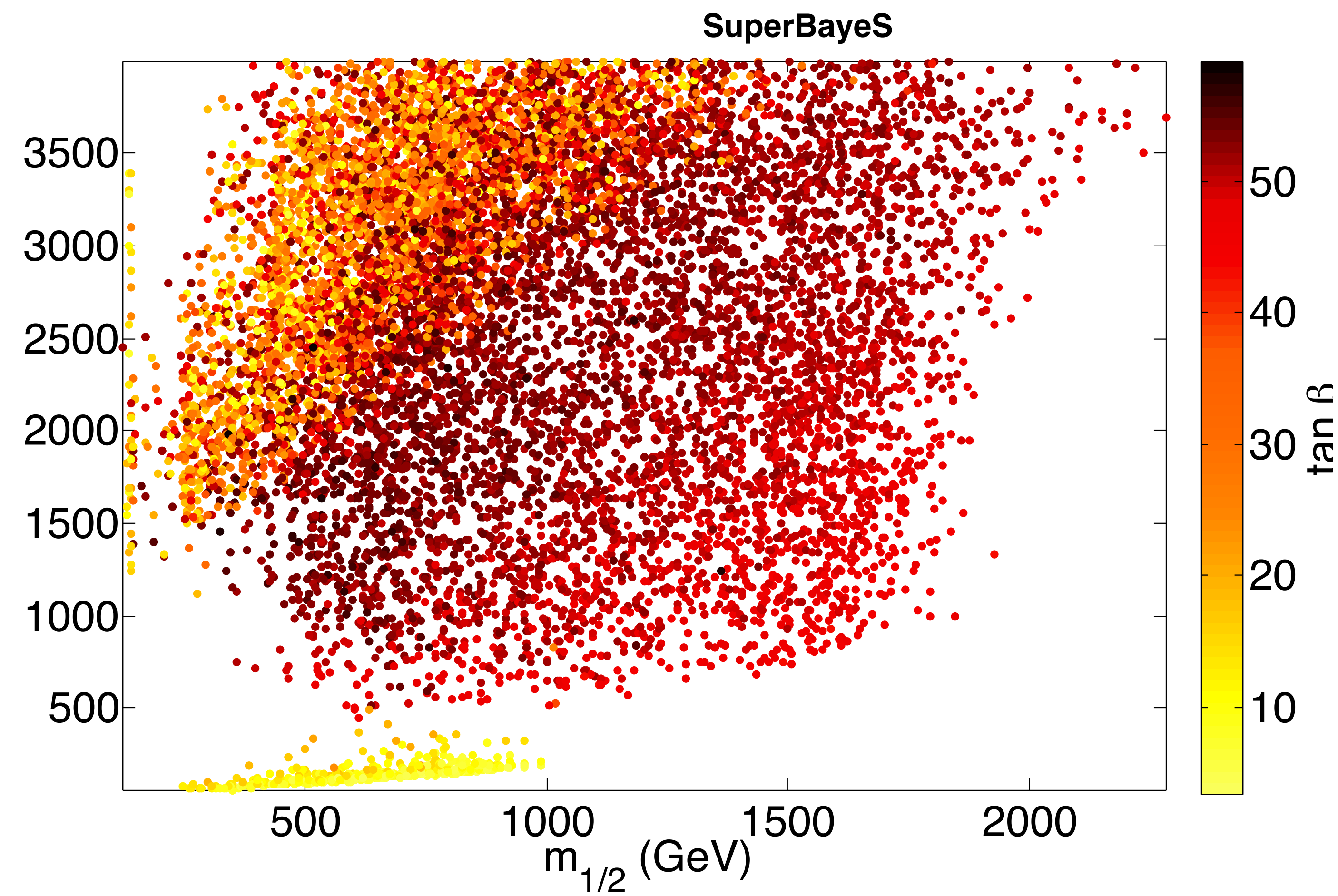


Constrained Minimal Supersymmetric Standard Model (4 parameters)
Strege, RT et al (2013)

Fancier stuff

Imperial College London





The simplest MCMC algorithm

Imperial College London

- Several (sophisticated) algorithms to build a MC are available: e.g. Metropolis-Hastings, Hamiltonian sampling, Gibbs sampling, rejection sampling, mixture sampling, slice sampling and more...

- Arguably the simplest algorithm is the **Metropolis (1954) algorithm:**

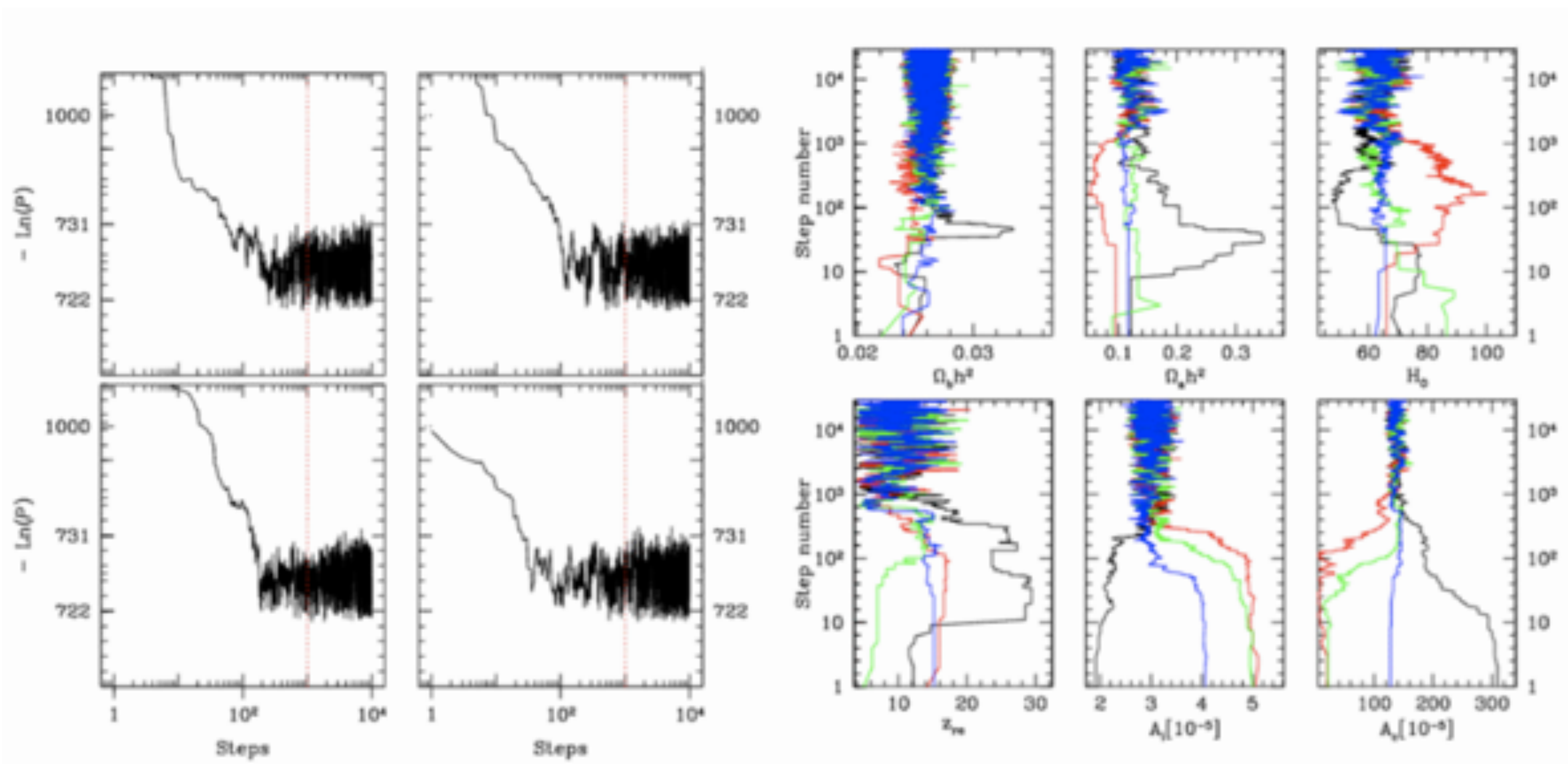
  - pick a starting location θ_0 in parameter space, compute $P_0 = p(\theta_0|d)$
  - pick a candidate new location θ_c according to a proposal density $q(\theta_0, \theta_1)$
  - evaluate $P_c = p(\theta_c|d)$ and accept θ_c with probability $\alpha = \min\left(\frac{P_c}{P_0}, 1\right)$
  - if the candidate is accepted, add it to the chain and move there; otherwise stay at θ_0 and count this point once more.

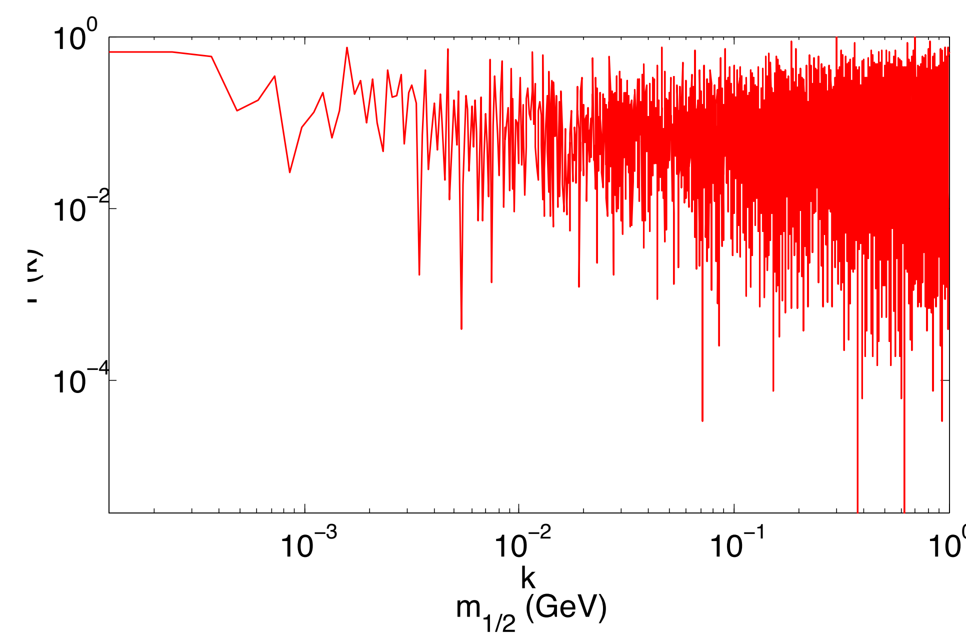
Practicalities

- Except for simple problems, achieving good MCMC **convergence** (i.e., sampling from the target) and **mixing** (i.e., all chains are seeing the whole of parameter space) can be tricky
- There are several diagnostics criteria around but none is fail-safe. Successful MCMC remains a bit of a black art!
- Things to watch out for:
  - Burn in time
  - Mixing
  - Samples auto-correlation

MCMC diagnostics

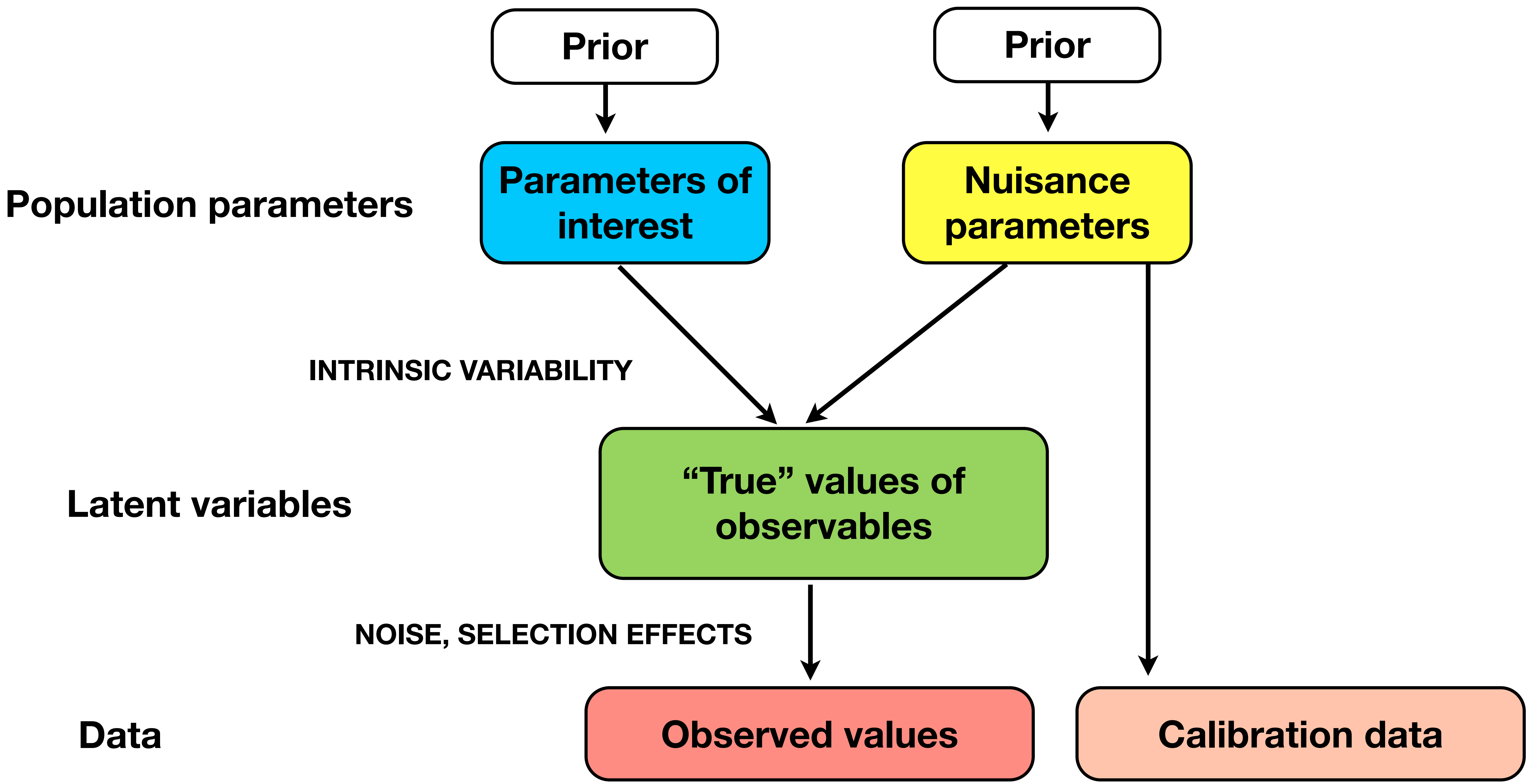
Burn in   Mixing   Power spectrum





(see astro-ph/0405462 for details)

Bayesian hierarchical models



At the heart of the method...

Imperial College London

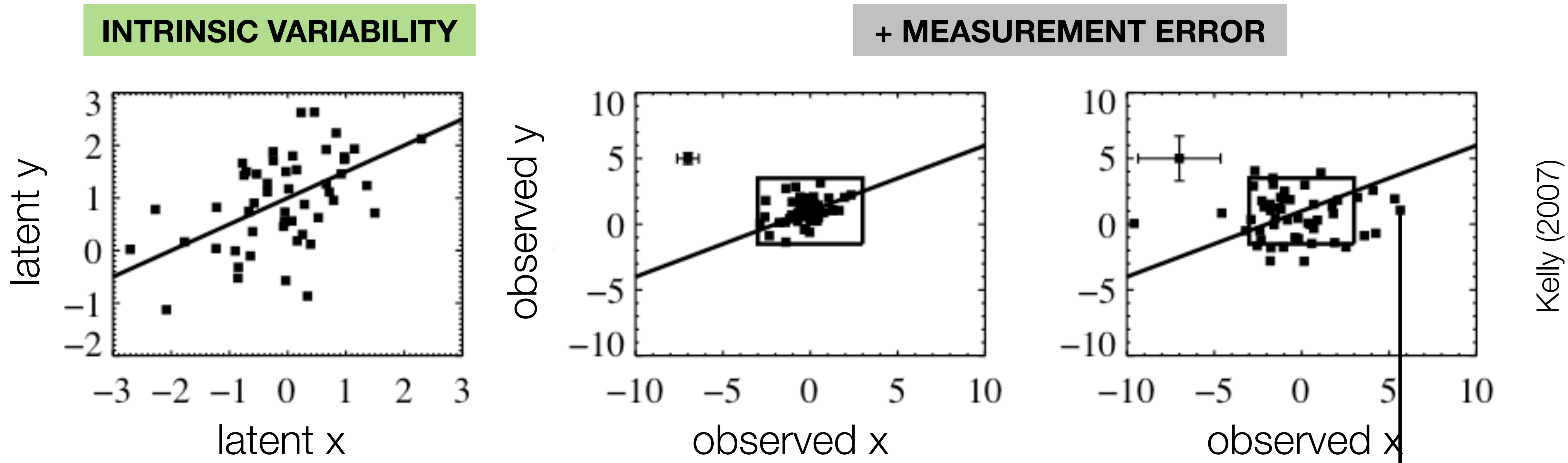
- ... lies the fundamental problem of **linear regression** in the presence of measurement errors on both the dependent and independent variable and intrinsic scatter in the relationship (e.g., Gull 1989, Gelman et al 2004, Kelly 2007):

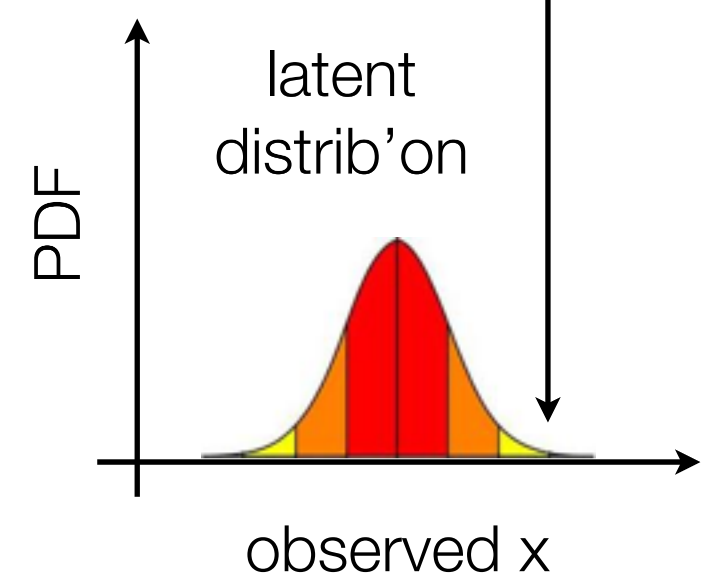
$$y_i = b + a x_i$$

$$x_i \sim p(x|\Psi) = \mathcal{N}_{x_i}(x_\star, R_x)$$ POPULATION DISTRIBUTION

$$y_i|x_i \sim \mathcal{N}_{y_i}(b + a x_i, \sigma^2)$$ INTRINSIC VARIABILITY

$$\hat{x}_i, \hat{y}_i | x_i, y_i \sim \mathcal{N}_{\hat{x}_i, \hat{y}_i}([x_i, y_i], \Sigma^2)$$ MEASUREMENT ERROR





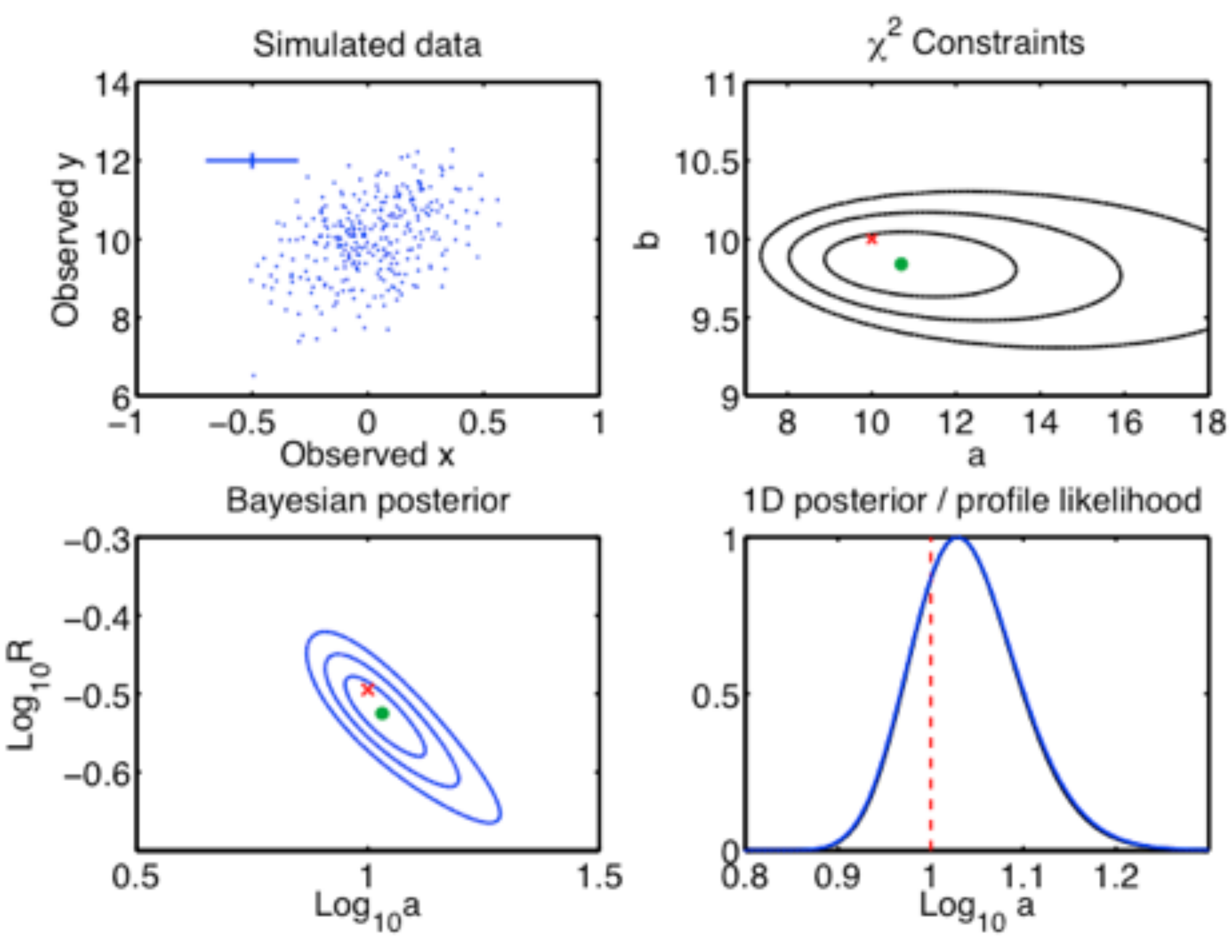
- Modeling the latent distribution of the independent variable accounts for "Malmquist bias"
- An observed x value far from the origin is more probable to arise from up-scattering (due to noise) of a lower latent x value than down-scattering of a higher (less probable) x value

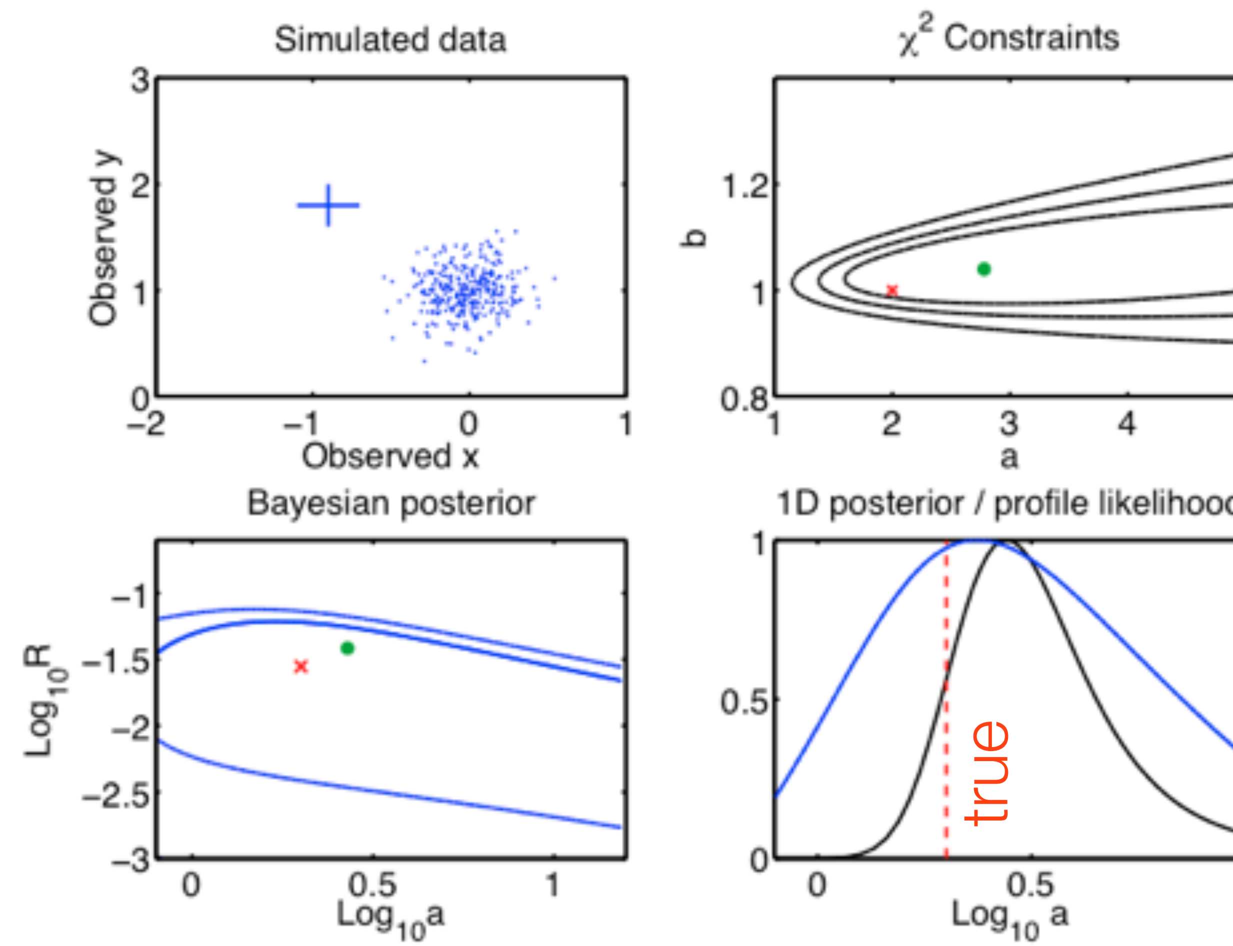
The key parameter is noise/population variance $\sigma_x \sigma_y / R_x$

$\sigma_x \sigma_y / R_x$ small

$$y_i = b + a x_i$$

$\sigma_x \sigma_y / R_x$ large





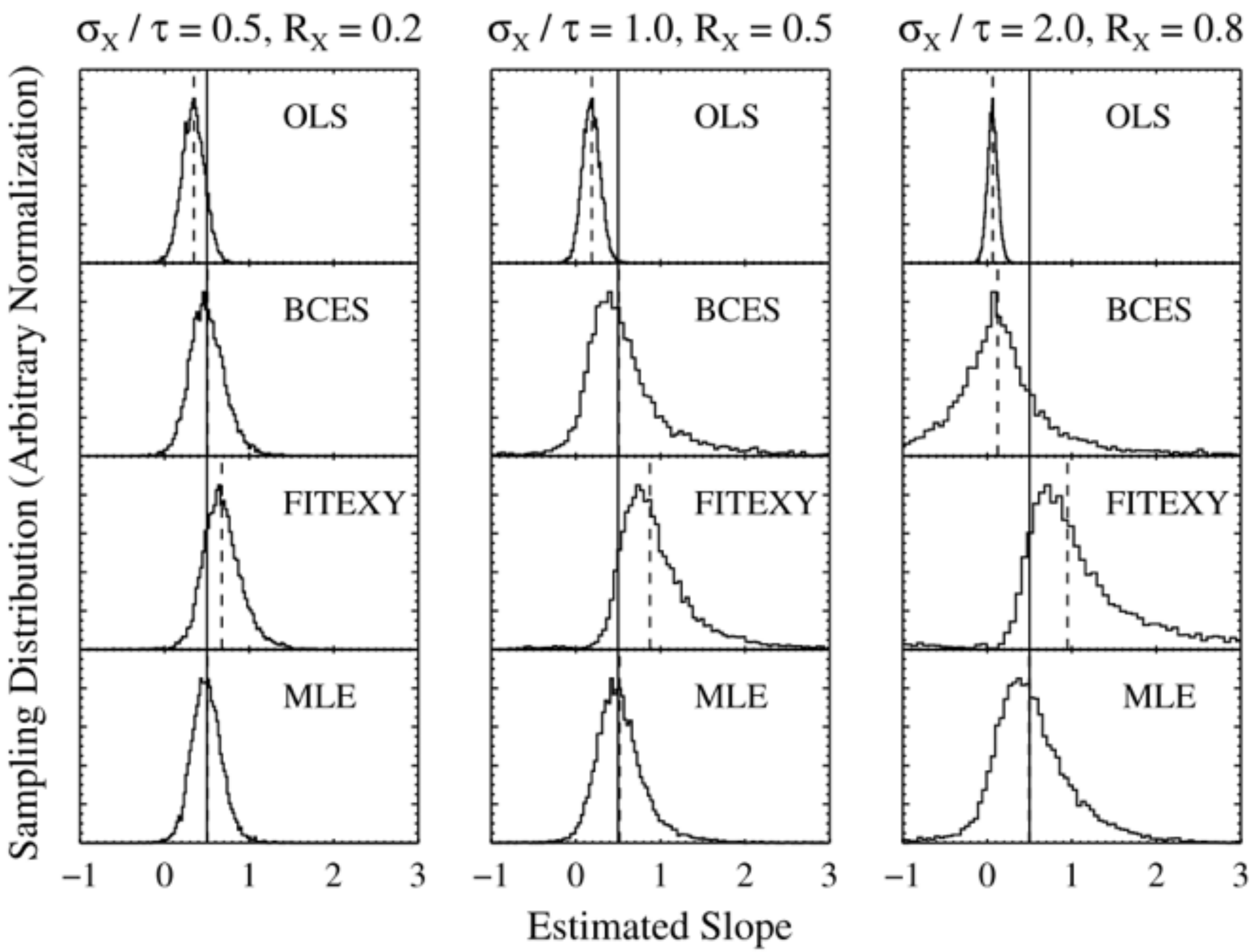
Bayesian marginal posterior identical to profile likelihood

Bayesian marginal posterior broader but less biased than profile likelihood

March, RT et al (2011)

Slope reconstruction

$R_x = \sigma_x^2/\mathrm{Var}(x)$: ratio of the covariate measurement variance to observed variance



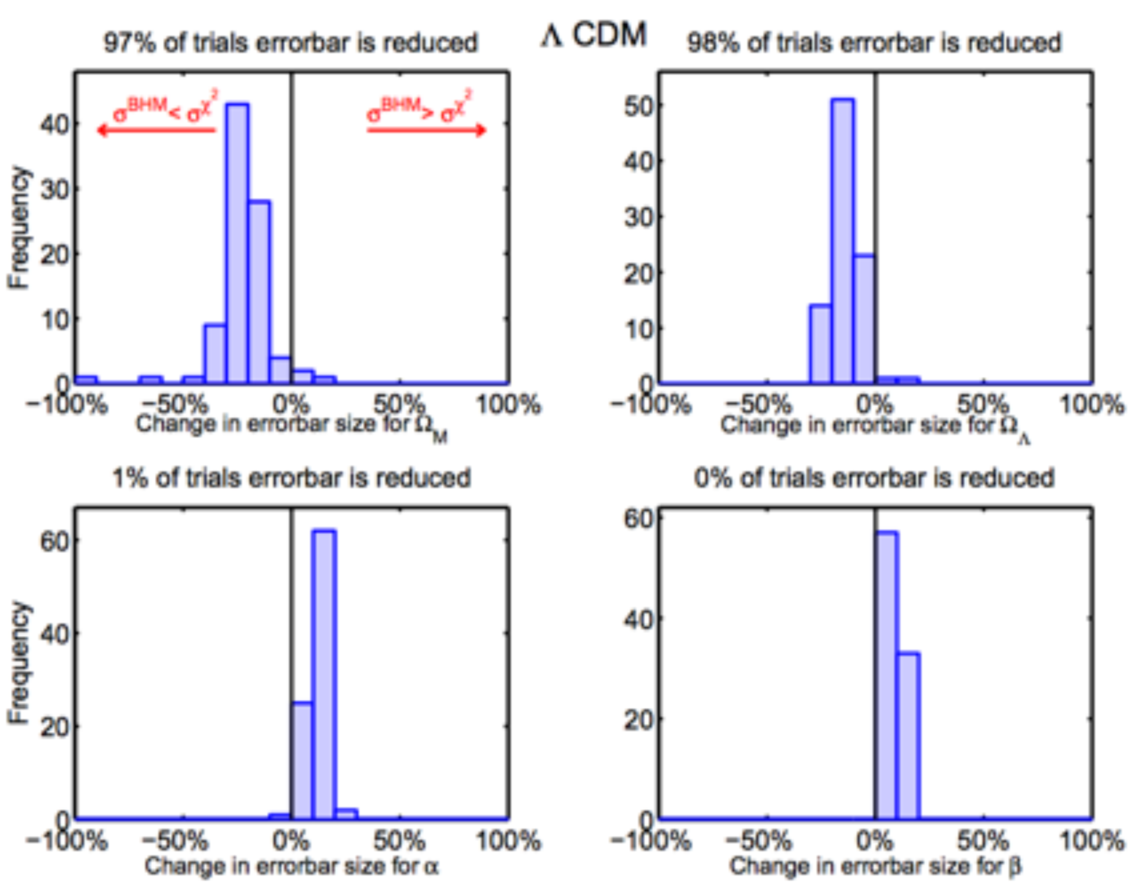
Kelly, Astr. J., 665, 1489-1506 (2007)

SNIa cosmology example

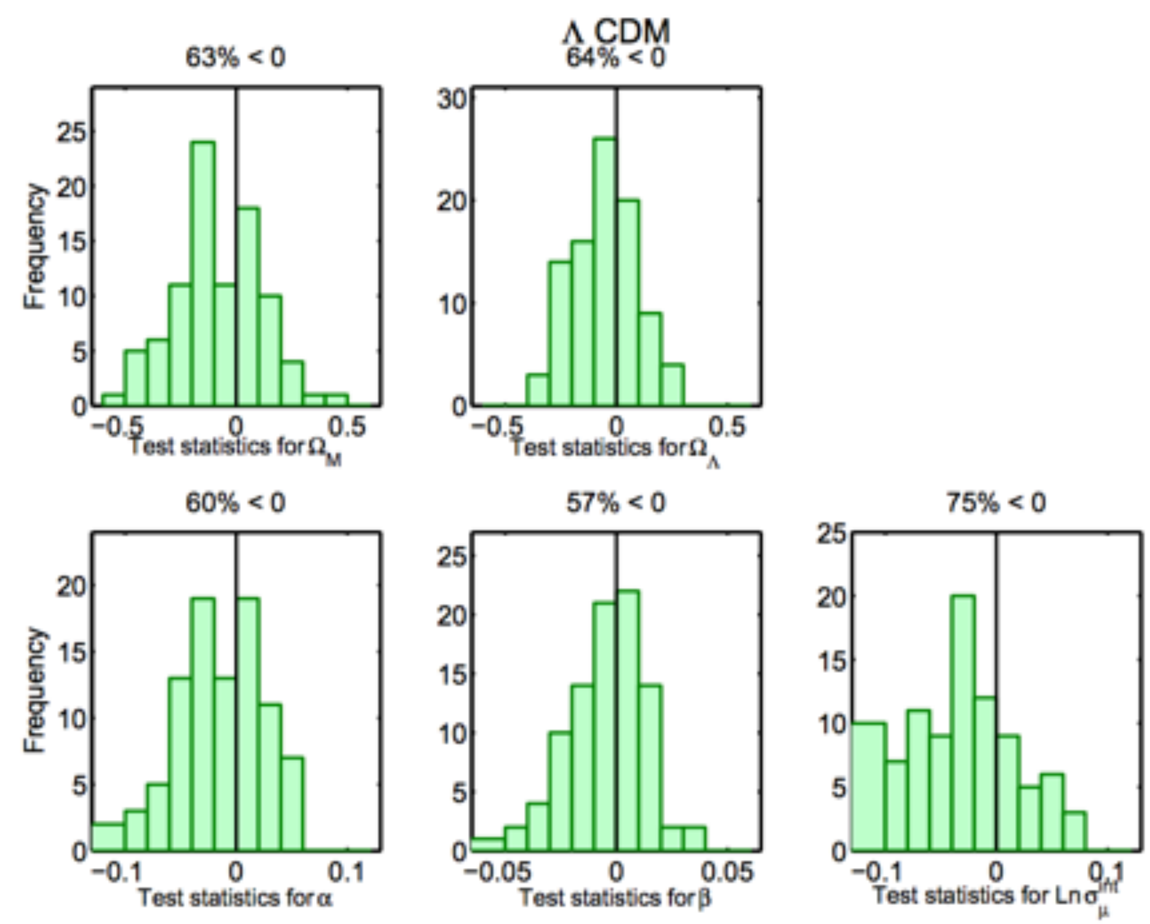
Imperial College London

- Comparing Bayesian Hierarchical approach to usual Chi-Square

Size of errorbars



Bias



March, RT et al, MNRAS 418(4),2308-2329 (2011)

Supernovae Type Ia Cosmology example

Imperial College London

- Coverage of Bayesian 1D marginal posterior CR and of 1D Chi[2] profile likelihood CI computed from 100 realizations

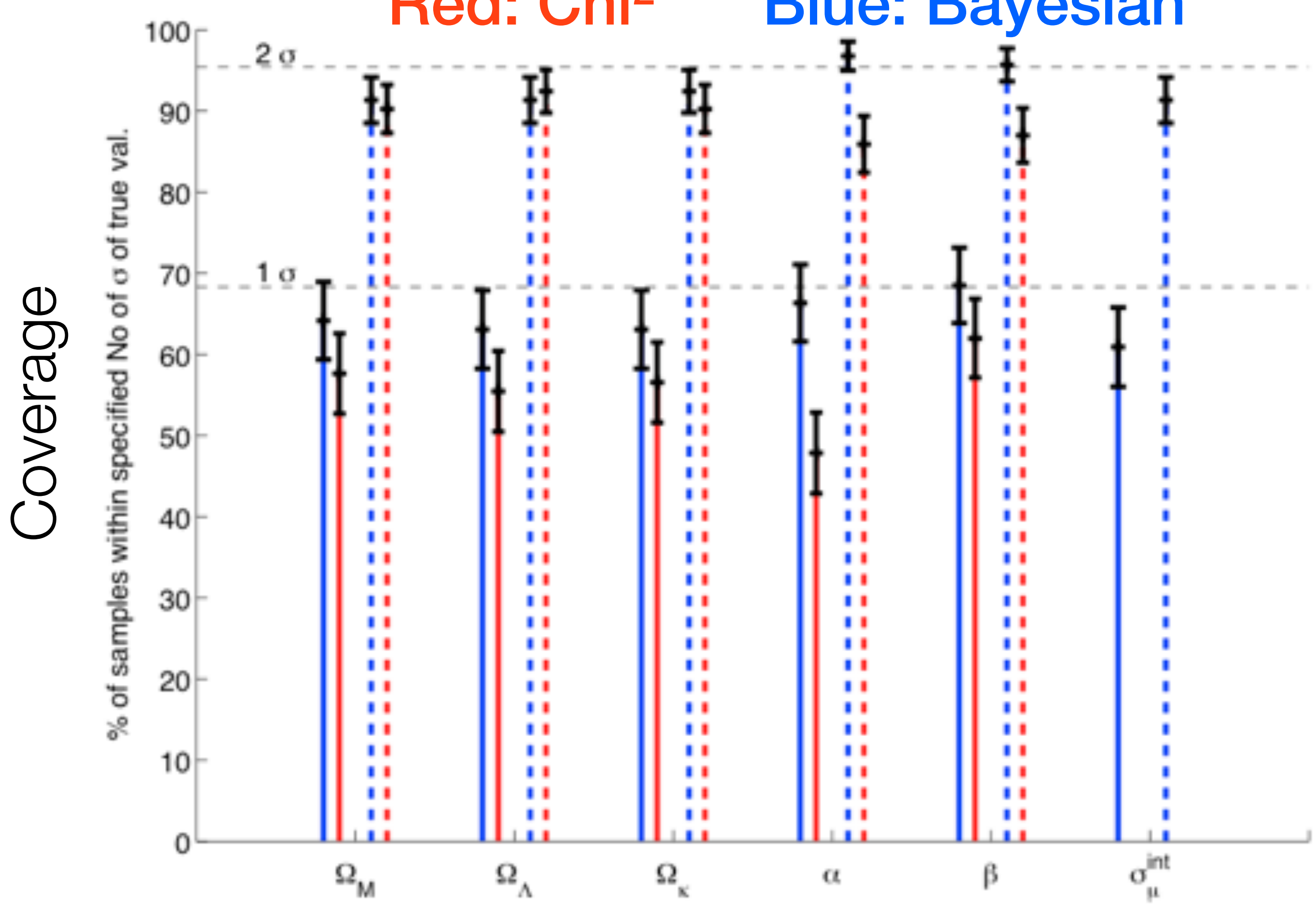
- Bias and mean squared error (MSE) defined as

$$\text{bias} = \langle \hat{\theta} - \theta_{\text{true}} \rangle$$

$$\text{MSE} = \text{bias}^2 + \text{Var}$$

$\hat{\theta}$ is the posterior mean (Bayesian) or the maximum likelihood value (Chi[2]).

Red: Chi[2] Blue: Bayesian



Results:

Coverage: generally improved (but still some undercoverage observed)

Bias: reduced by a factor ~ 2-3 for most parameters

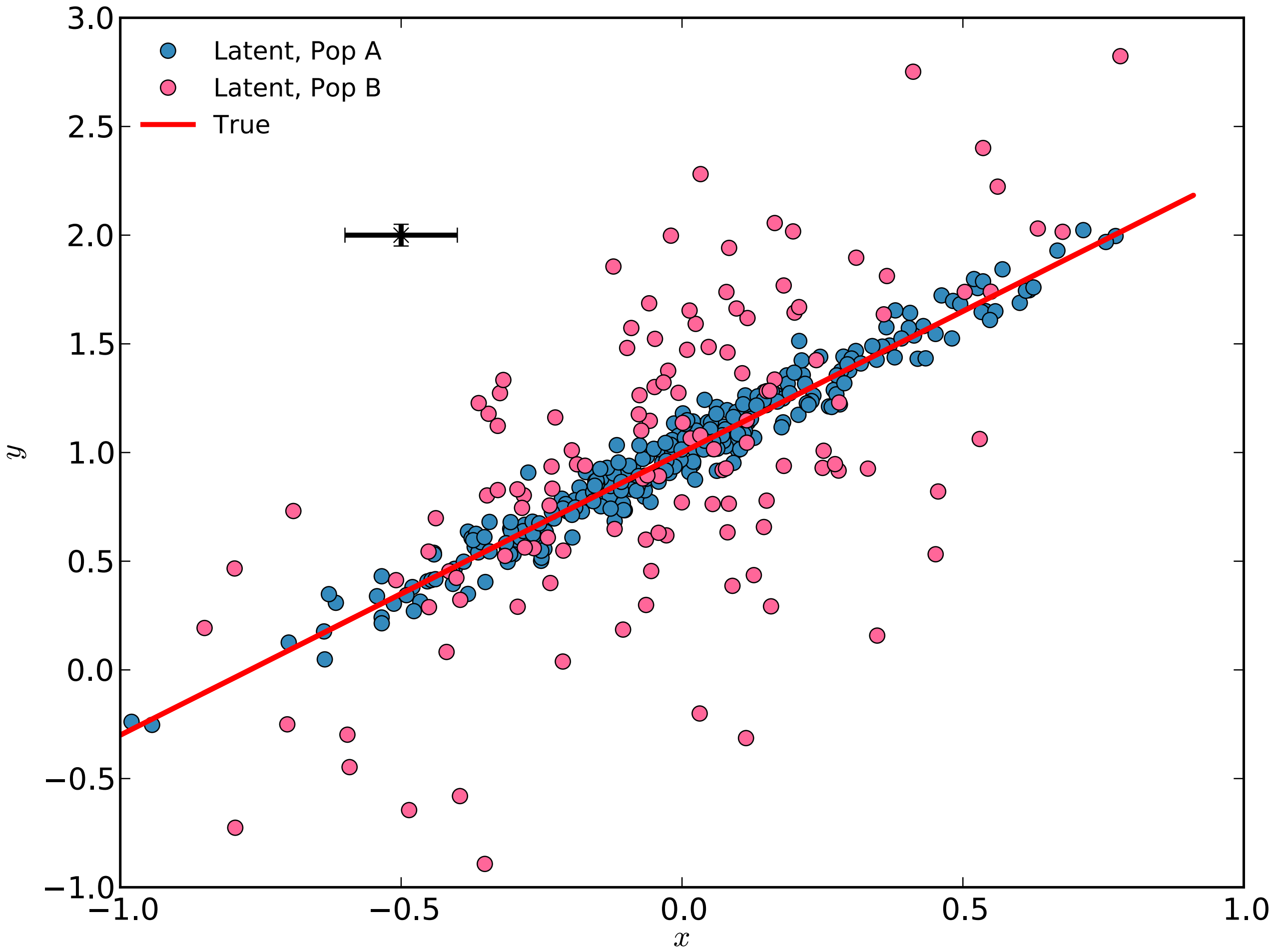
MSE: reduced by a factor 1.5-3.0 for all parameters

Adding object-by-object classification

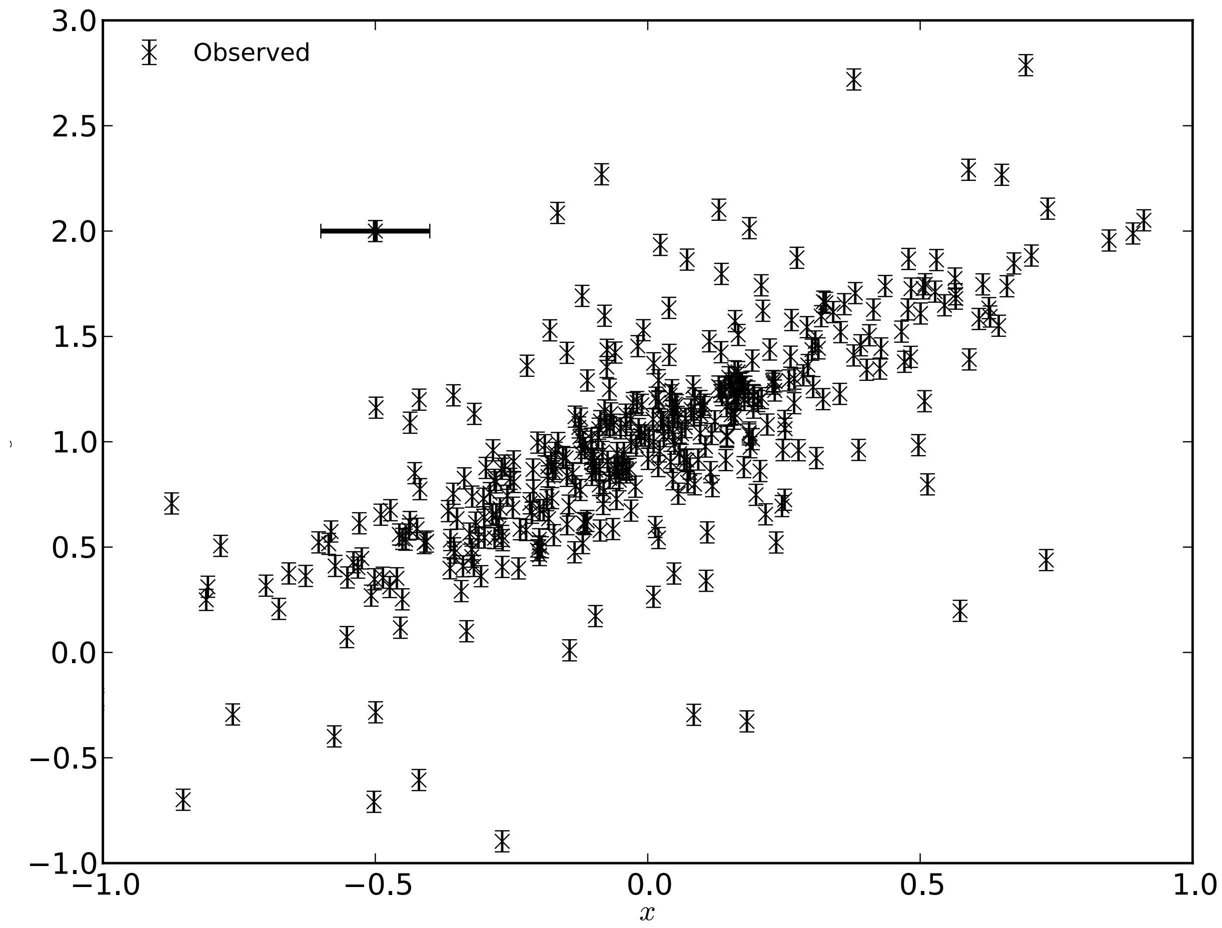
Imperial College London

- "Events" come from two different populations (with different intrinsic scatter around the same linear model), but we ignore which is which:

LATENT



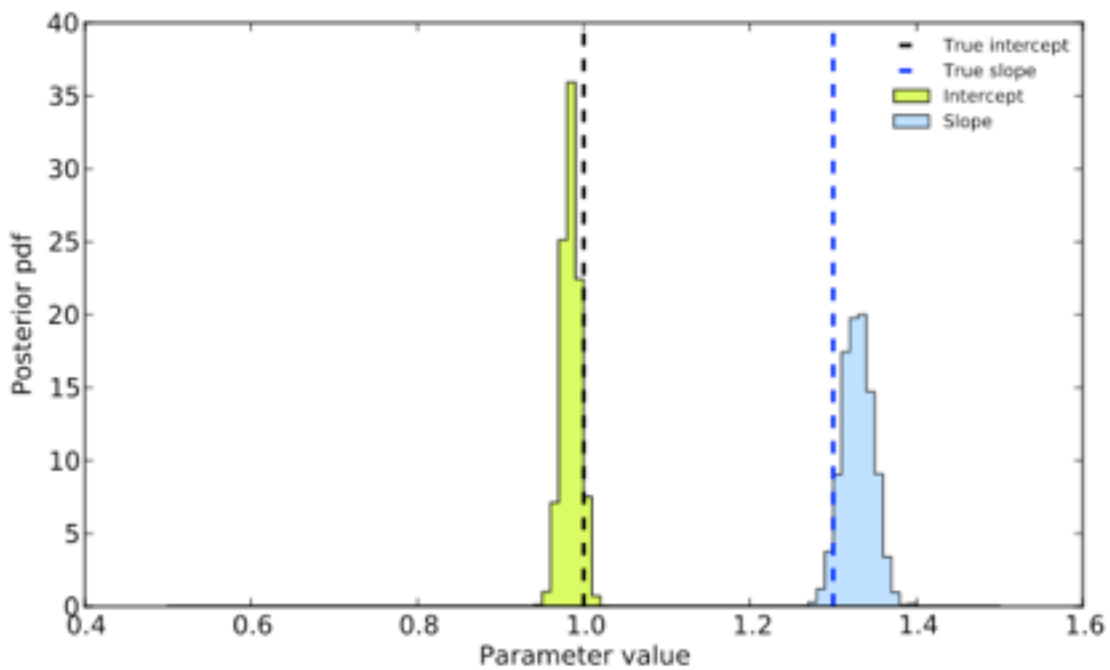
OBSERVED

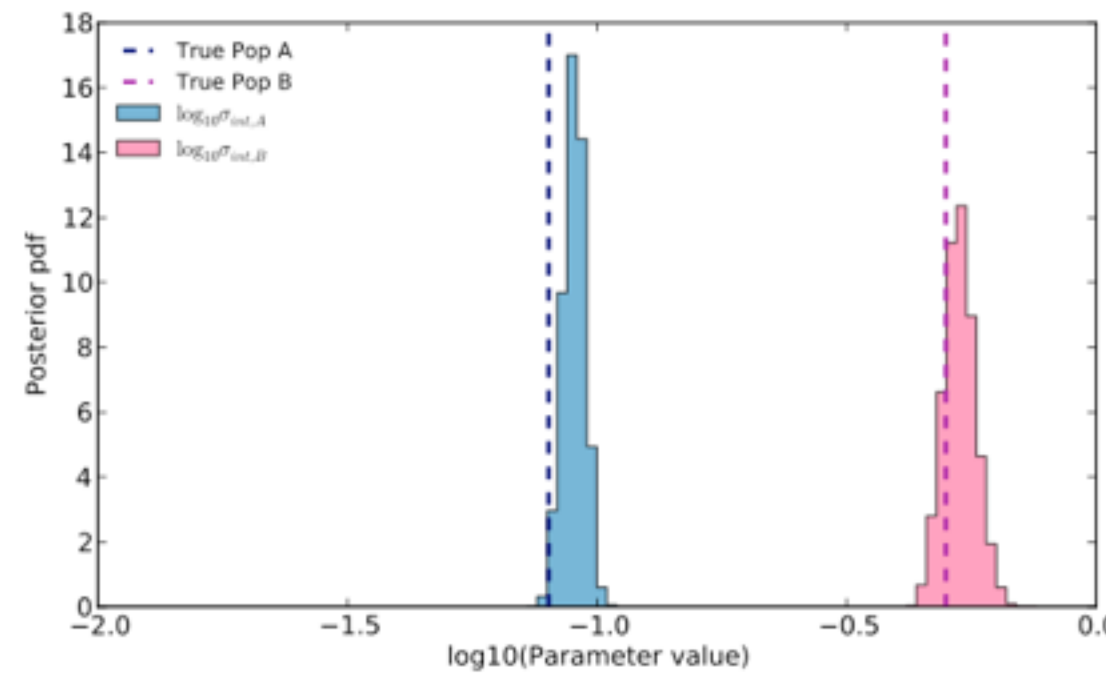


ICIC

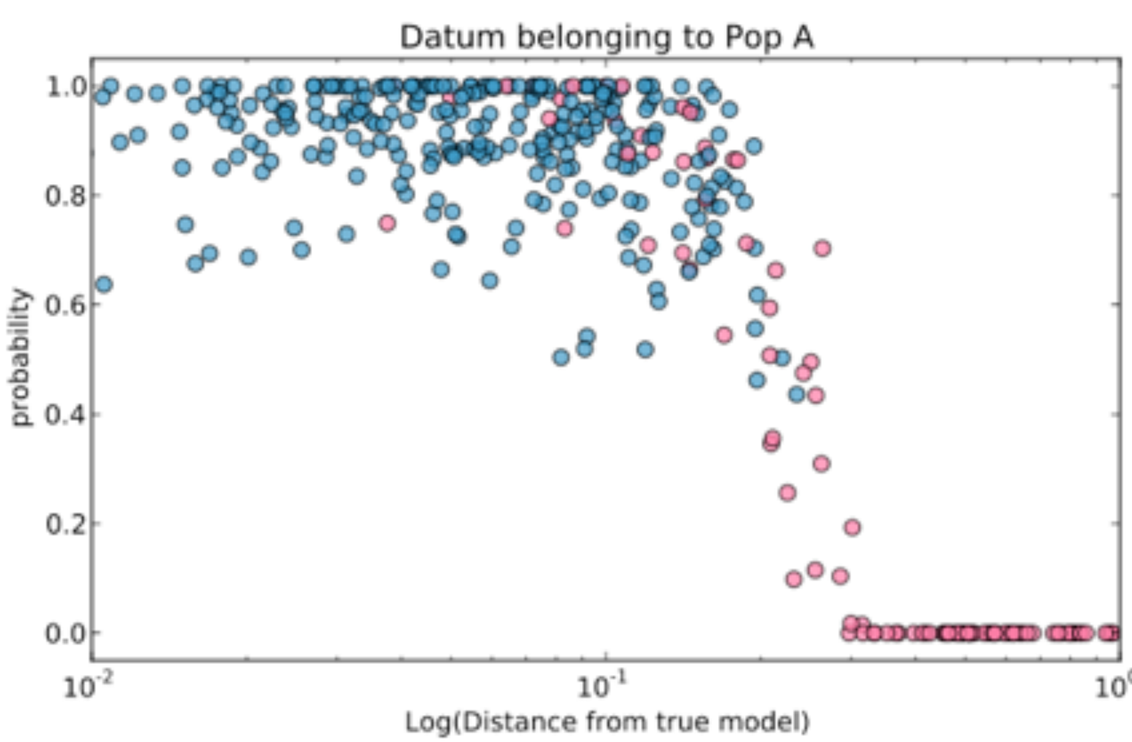
Reconstruction (N=400)

Parameters of interest

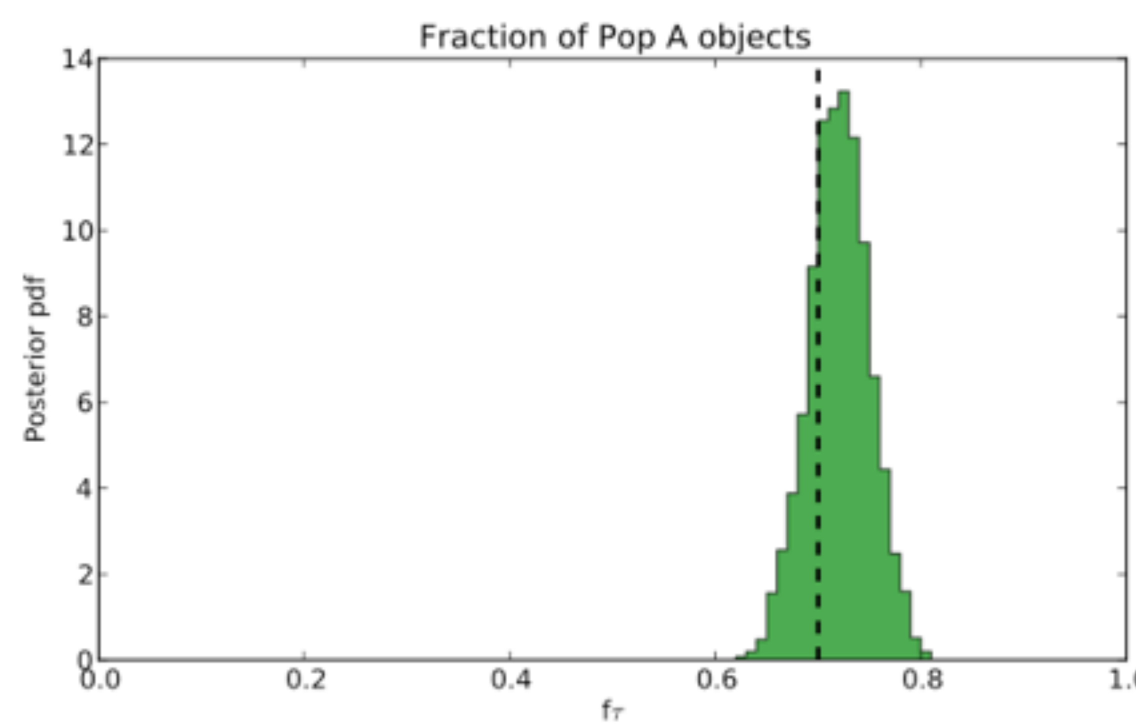




Classification of objects



Population-level properties



Prediction and optimization

The Bayesian perspective

- In the Bayesian framework, we can use present-day knowledge to produce probabilistic forecasts for the outcome of a future measurement
- This is **not** limited to assuming a model/parameter value to be true and to determine future errors
- Many questions of interest today are of model comparison: e.g.
  - is dark energy Lambda or modified gravity?
  - is dark energy evolving with time?
  - Is the Universe flat or not?
  - Is the spectrum of perturbations scale invariant or not?

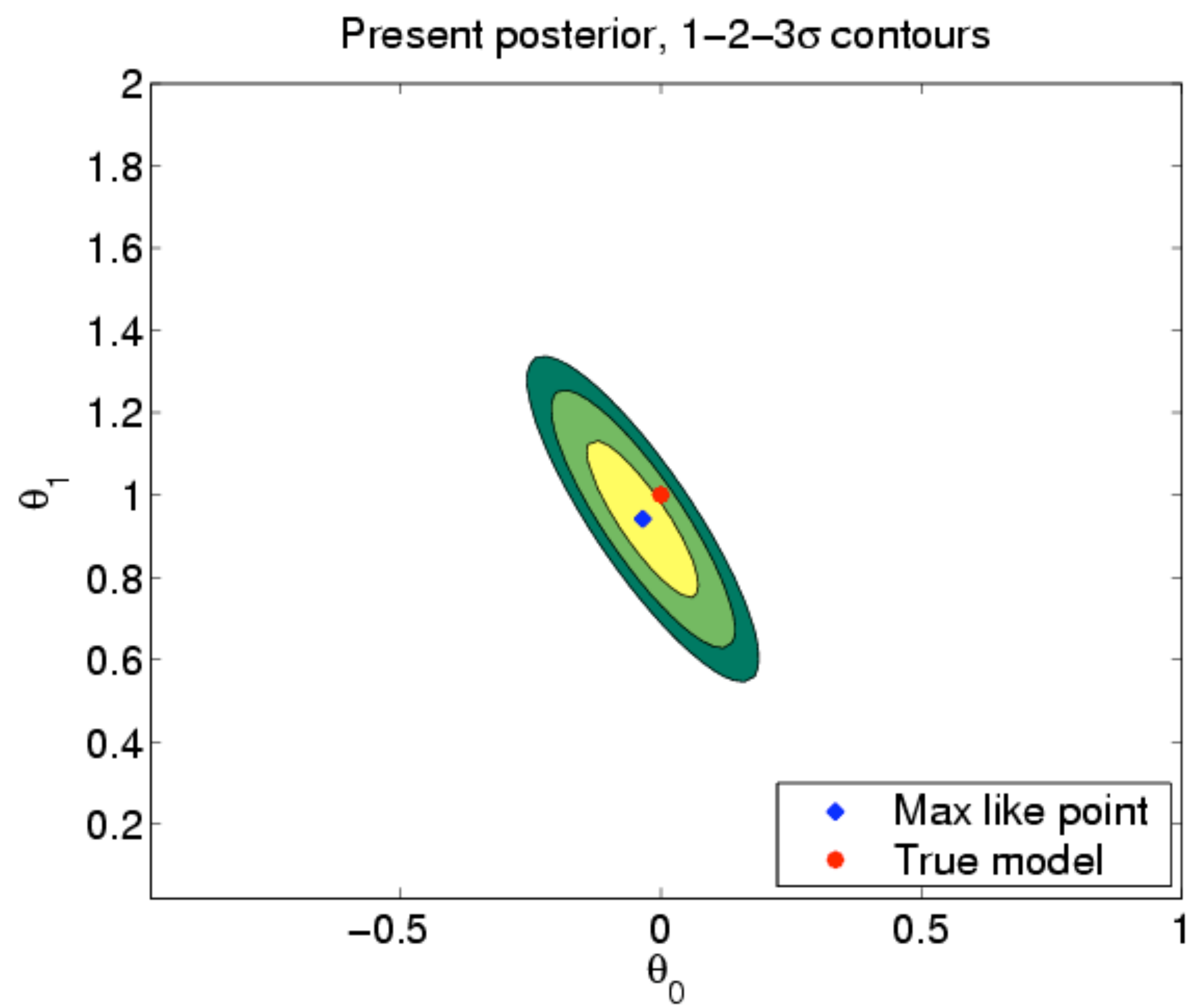
Predictions for future observations

- Toy model: the linear Gaussian model (see Exercices 7-9)

  $y = \theta_0 + x\theta_1$
  $y - Fx = \varepsilon$

- Gaussian noise on ε

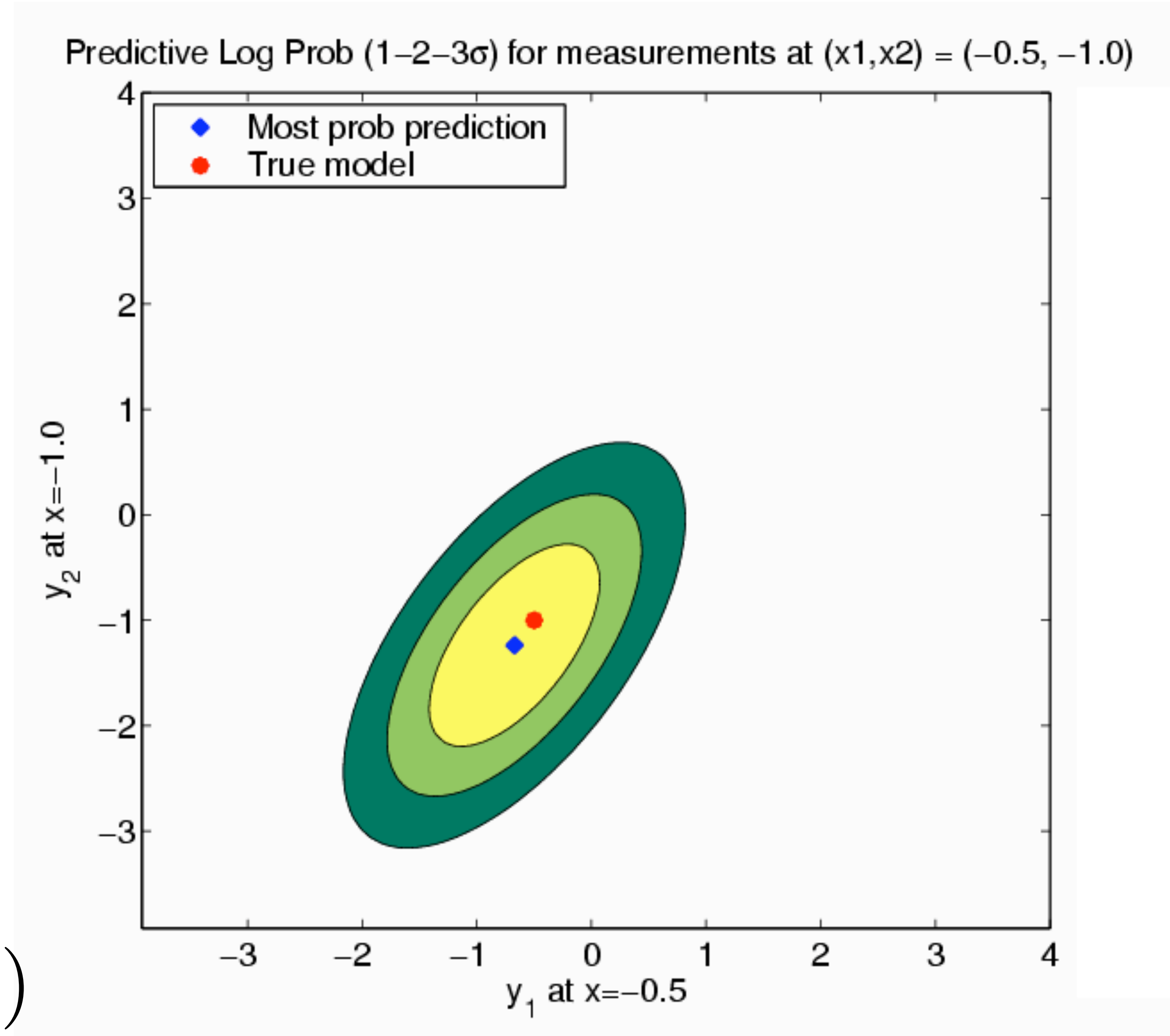
- True values: $(\theta_0, \theta_1) = (0,1)$



The predictive distribution

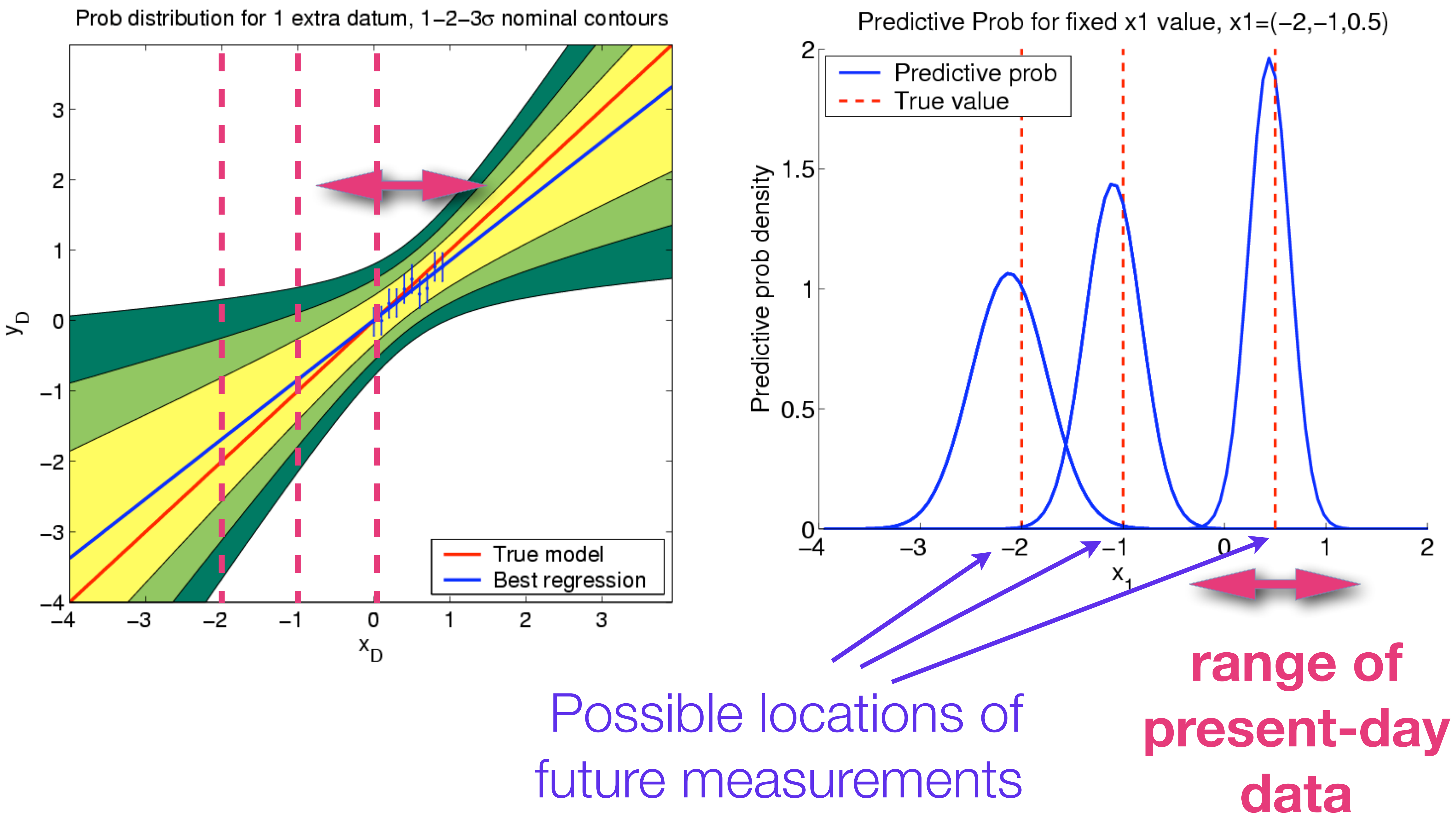
- Use present knowledge (**and uncertainty!)** to predict what a future measurement will find (with corresponding probability)
- True values: $(\theta_0, \theta_1) = (0,1)$
- Present-day data: d
- Future data: D

$$P(D|d) = \int d\theta P(D|\theta) P(\theta|d)$$



Predictive probability = future likelihood weighted by present posterior

Predictive distribution



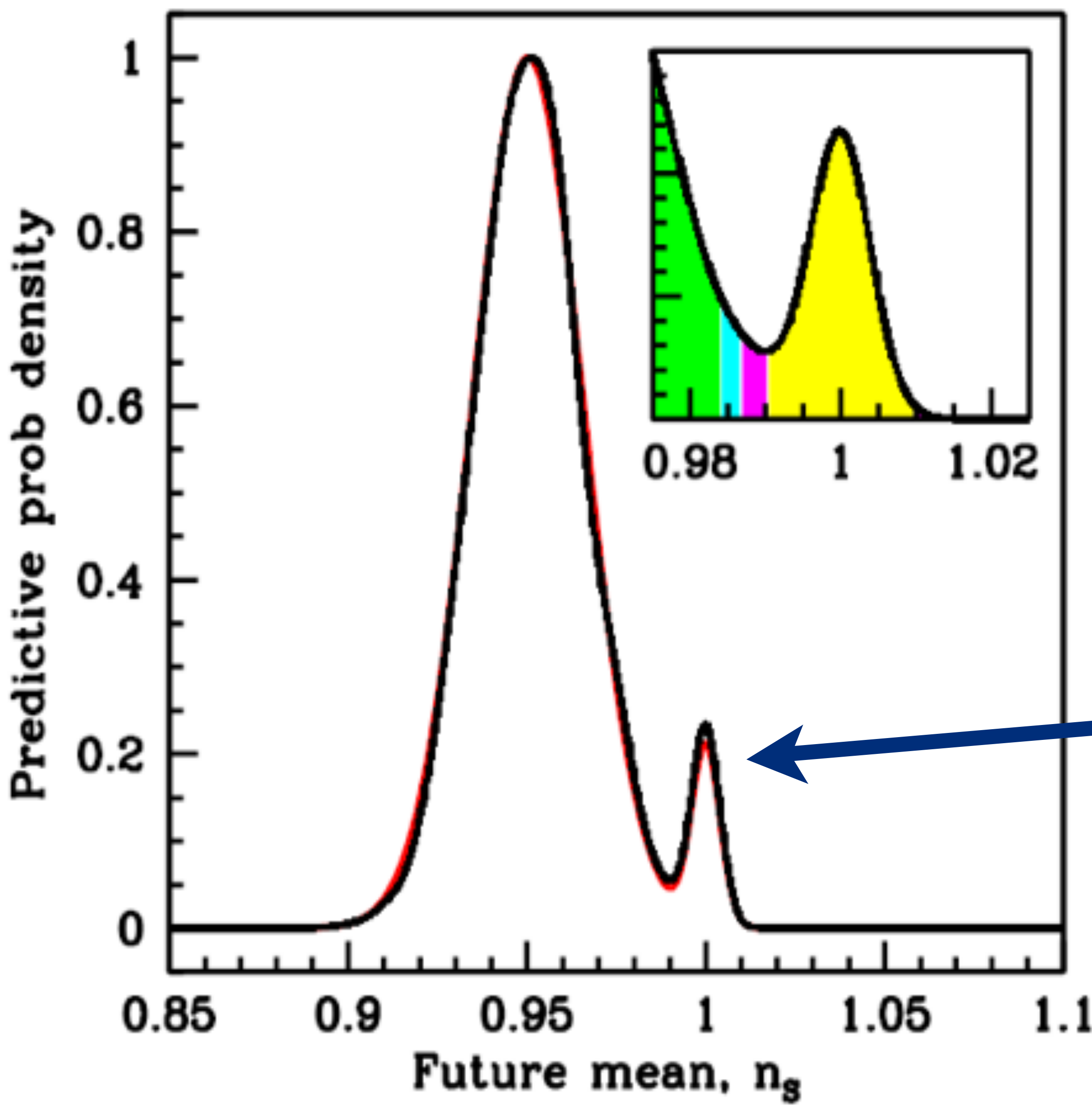
Extending the power of forecasts

Imperial College London

- Thanks to predictive probabilities we can increase the scope and power of forecasts:
- **Level 0:** assume a model M and a fiducial value for the parameters, θ* produce a forecast for the errors that a future experiment will find *if M and θ* are the correct choices*
- **Level 1:** average over current parameter uncertainty within M
- **Level 2:** average over current model uncertainty: replace M by M_1, M_2,...

Predictive posterior odds distribution

Bayes factor forecast for Planck



$n_s=1$ vs $0.8<n_s<1.2$

P(lnB < -5) = 0.93
P(-5<lnB<0) = 0.01
P(lnB > 0) = 0.06

Model uncertainty
$P(n_s=1|\text{WMAP3+}) = 0.05$

Trotta (2008), Parkinson et al (2006), Pahud et al (2006)

Experiment design

Utility and optimization

Imperial College London

- The optimization problem is fully specified once we define a **utility function U** depending on the outcome *e* of a future observation (e.g., scientific return). We write for the utility U(e, o, θ), where *o* is the current experiment and θ are the true values of the parameters of interest
- We can then evaluate the **expected utility:**

$$\mathcal{E}[U|e,o] = \int d\theta U(\theta,e,o)P(\theta|o)$$

Example: an astronomer measures y = θ x (with Gaussian noise) at a few points 0 < x < 1. She then has a choice between building 2 equally expensive instruments to perform a new measurement:

1. Instrument (e) is as accurate as today's experiments but extends to much larger values of x (to a maximum x_{max})
2. Instrument (a) is much more accurate but it is built in such a way as has to have a "sweet spot" at a certain value of y, call it y*, and much less accurate elsewhere

Which instrument should she go for?

ICIC

Answer

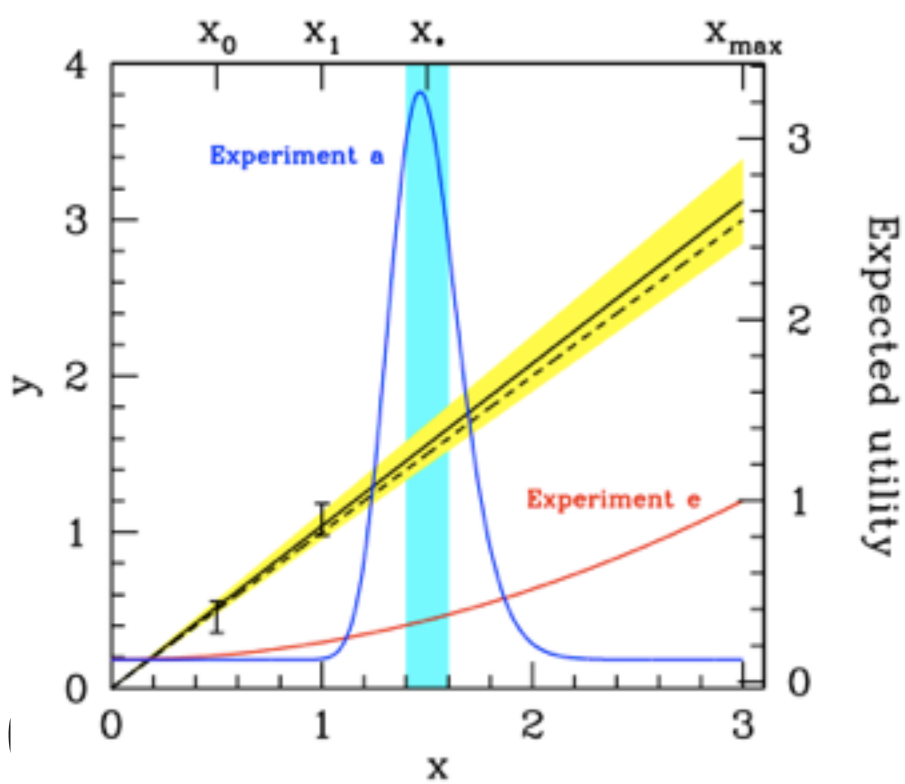
Imperial College London

- The answer depends on how good her current knowledge is - i.e. is the current uncertainty on θ* small enough to allow her to target accurately enough x=x* so that she can get to the "sweet spot" y*= θ*x*?
  (try it out for yourself! *Hint: use for the utility the inverse variance of the future posterior on θ and assume for the noise levels of experiment (a) the toy model:*

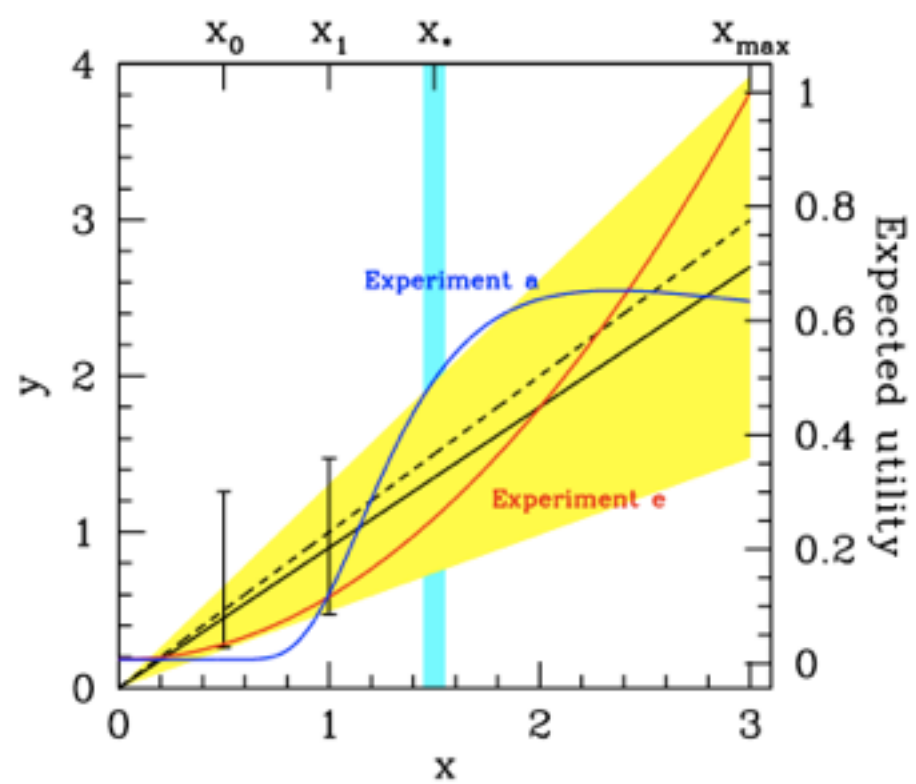
$$\tau_a^2 = \tau_*^2 \exp\left(\frac{(y-y_\star)^2}{2\Delta^2}\right)$$

where y is the location of the sweet spot and Δ is the width of the sweet spot)*

Small uncertainty



Large uncertainty



Trotta et al, in Bayesian methods in cosmology (CUP, 2010)

ICI

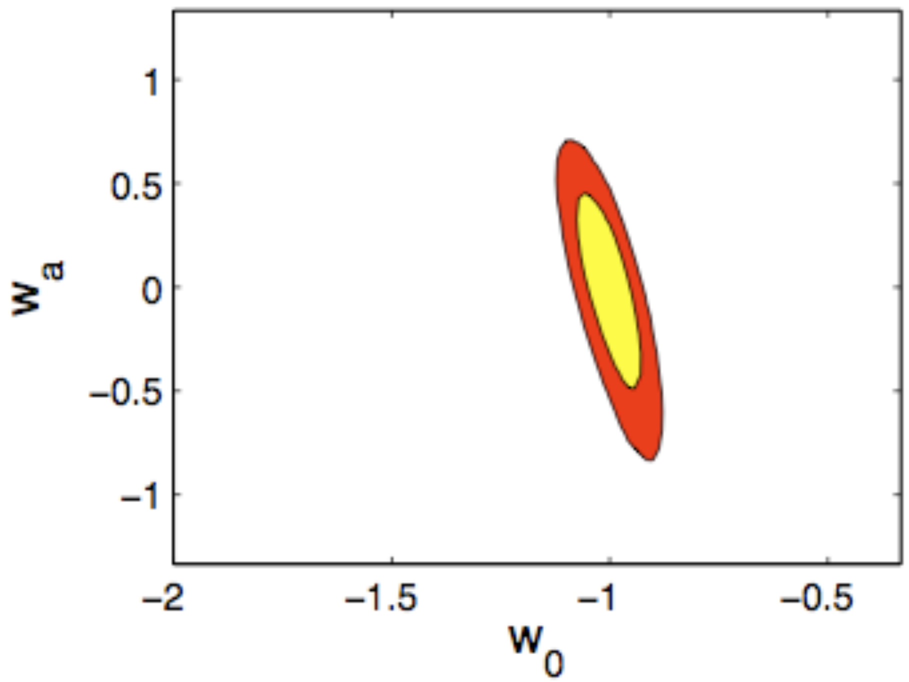
Making predictions: Dark Energy

Imperial College London

A model comparison question: is dark energy Lambda, i.e. $(w_0, w_a) = (-1, 0)$?
How well will the future probe SNAP be able to answer this?

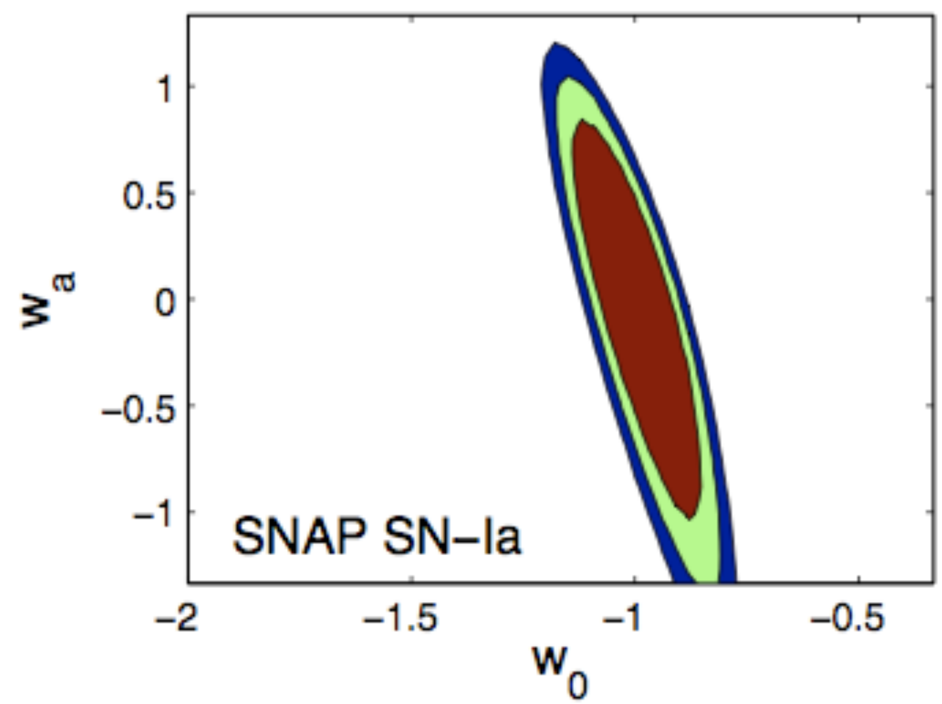
Mukherjee et al (2006)

Fisher Matrix



Simulates from LCDM
Assumes LCDM is true
Ellipse not invariant when changing model assumptios

Bayesian evidence



Simulate from all DE models
Assess "model confusion"
Allows to discriminate against LCDM

Key points

Imperial College London

- Predictive distributions incorporate present uncertainty in forecasts for the future scientific return of an experiment
- Experiment optimization requires the specification of an utility function. The “best” experiment is the one that maximises the expected utility.