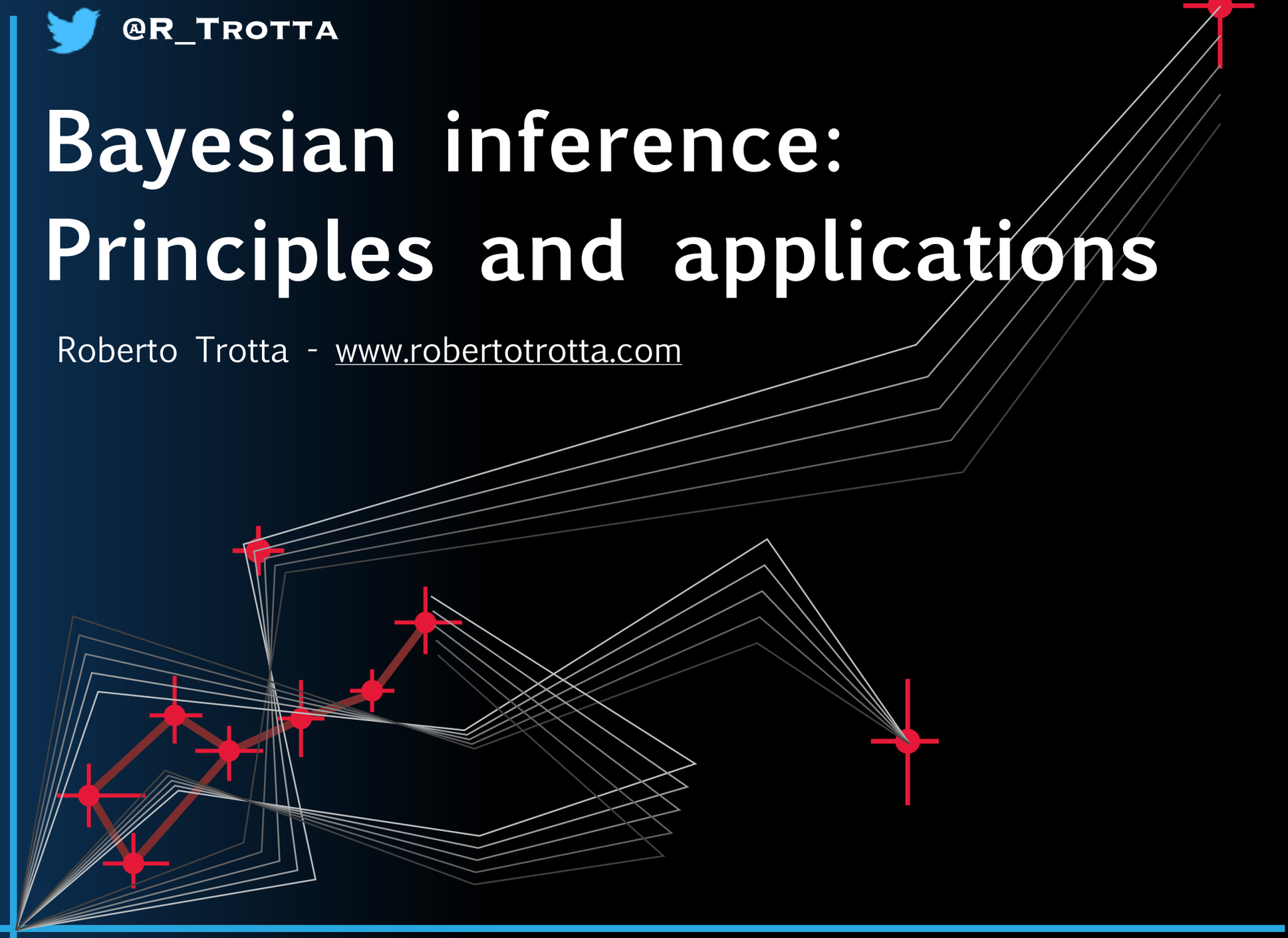
@R_Trotta

Bayesian inference: Principles and applications

Roberto Trotta - www.robertotrotta.com



ADA IX Summer School
Valencia, May 2018

ICIC
Imperial Centre
for Inference & Cosmology

Imperial College
London

Introductions

Imperial College London

- Your research interests in ~3 words (not a sentence!)
- How data intensive is your work? Range 1-5 (1 = very little, 5 = a lot)
- Your name and Institution

Bayes Theorem

Imperial College London

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

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

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

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

The equation of knowledge

Imperial College London

Consider two propositions A, B.

A = it will rain tomorrow, B = the sky is cloudy

A = the Universe is flat, B = observed CMB temperature map

Bayes' Theorem

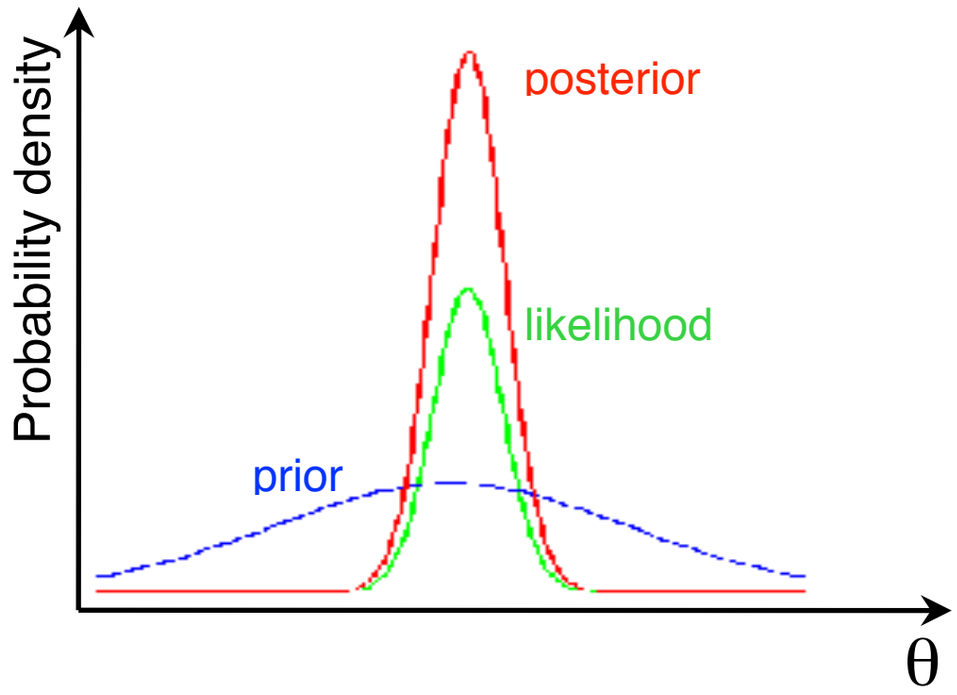
$$P(A|B)P(B) = P(A,B) = P(B|A)P(A)$$

Replace A → **θ** (the parameters of model **M**) and B → **d** (the data):

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

state of knowledge after | **information from the data** | **state of knowledge before**

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$



Why does Bayes matter?

Imperial College London

This is what our scientific questions are about (the posterior)

This is what classical statistics is stuck with (the likelihood)

$$P(\text{hypothesis}|\text{data}) \neq P(\text{data}|\text{hypothesis})$$

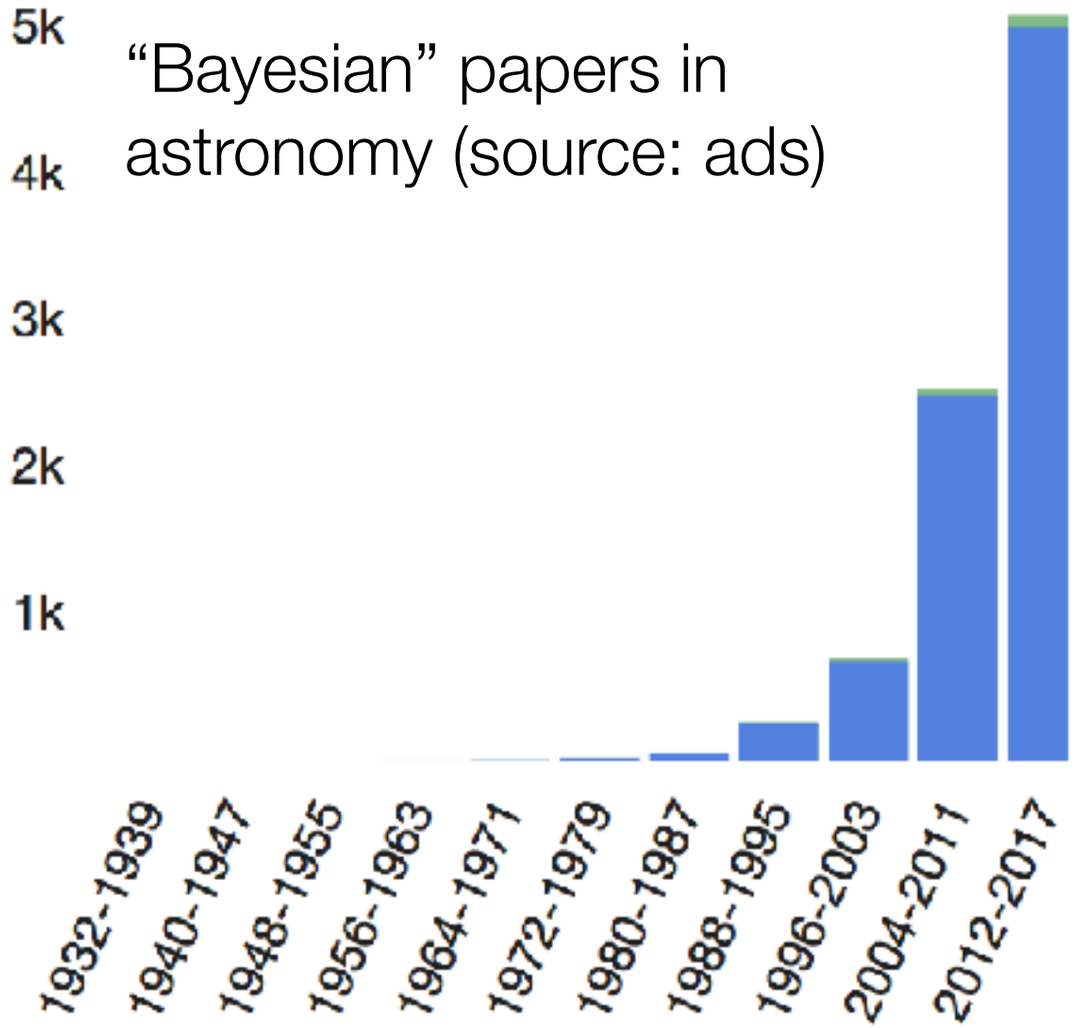
Example: is a randomly selected person female? (Hypothesis)

Data: the person is pregnant (d = pregnant)

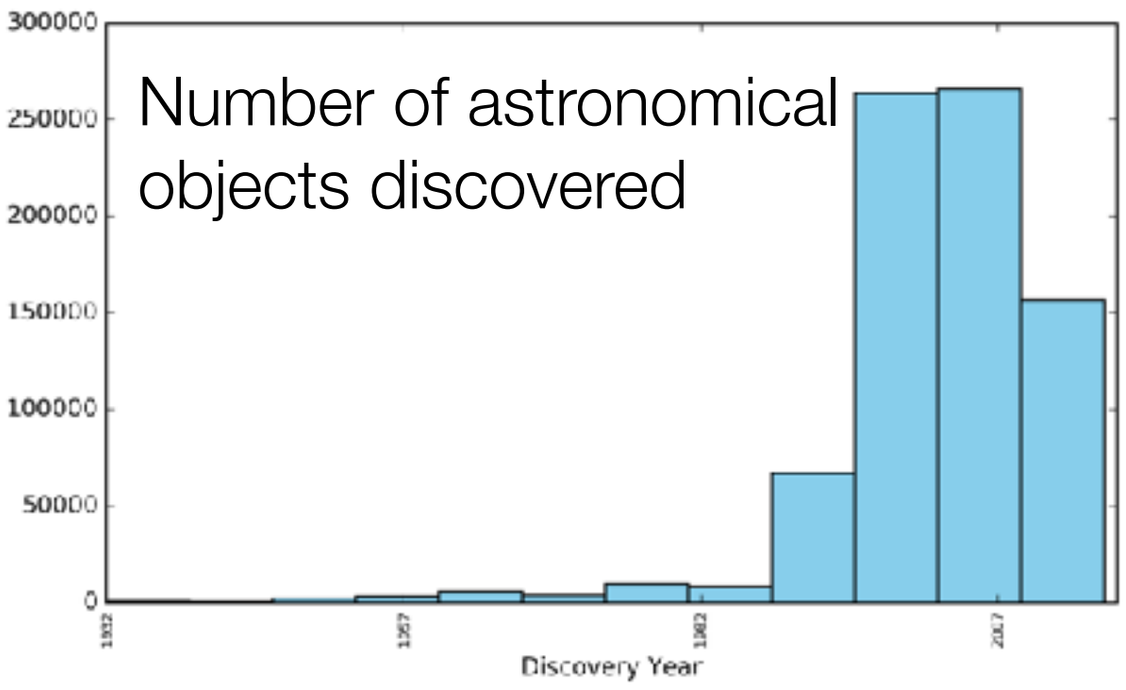
$$P(\text{female} \mid \text{pregnant}) = 1 \qquad P(\text{pregnant} \mid \text{female}) = 0.03$$

"Bayesians address the question everyone is interested in by using assumptions no–one believes, while frequentists use impeccable logic to deal with an issue of no interest to anyone"

Louis Lyons

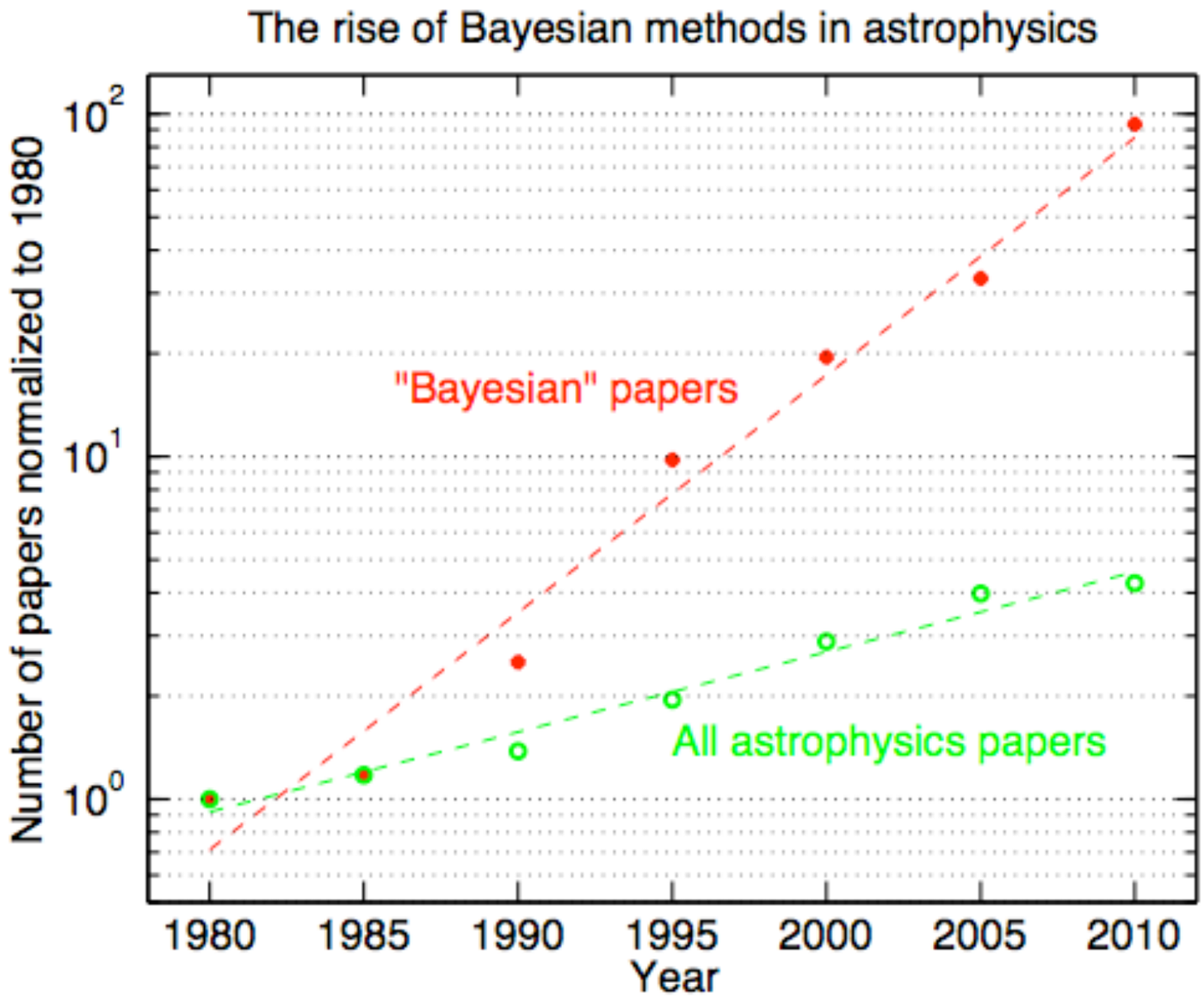


2000s: The age of (Bayesian) astrostatistics



Bayesian methods on the rise

Imperial College London



The real reasons to be Bayesian...

Imperial College London

... because it works!

- **Efficiency:** exploration of high-dimensional parameter spaces (e.g. with appropriate Markov Chain Monte Carlo) scales approximately linearly with dimensionality.

- **Consistency:** uninteresting (but important) parameters (e.g., instrumental calibration, unknown backgrounds) can be integrated out from the posterior with almost no extra effort and their uncertainty propagated to the parameters of interest.

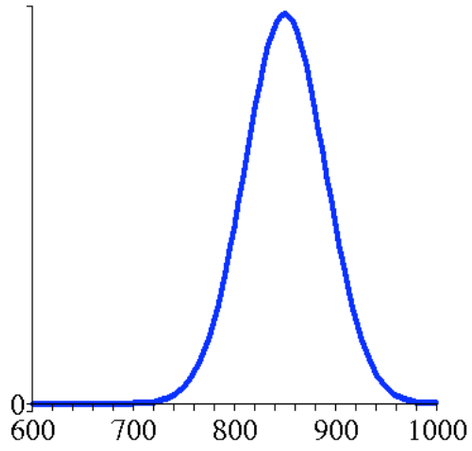
- **Insight:** having to define a prior forces the user to think about their assumptions! Whenever the posterior is strongly dependent on them, this means the data are not as constraining as one thought. “There is no inference without assumptions”.

The matter with priors

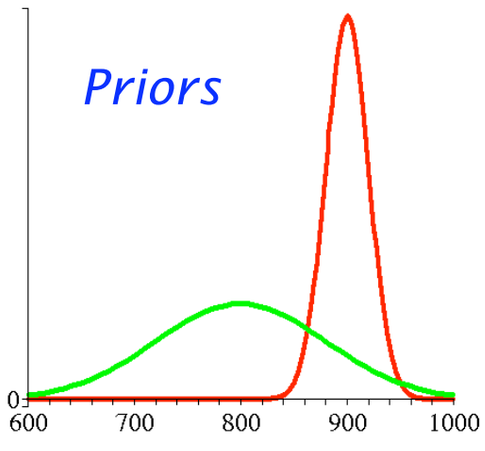
Imperial College London

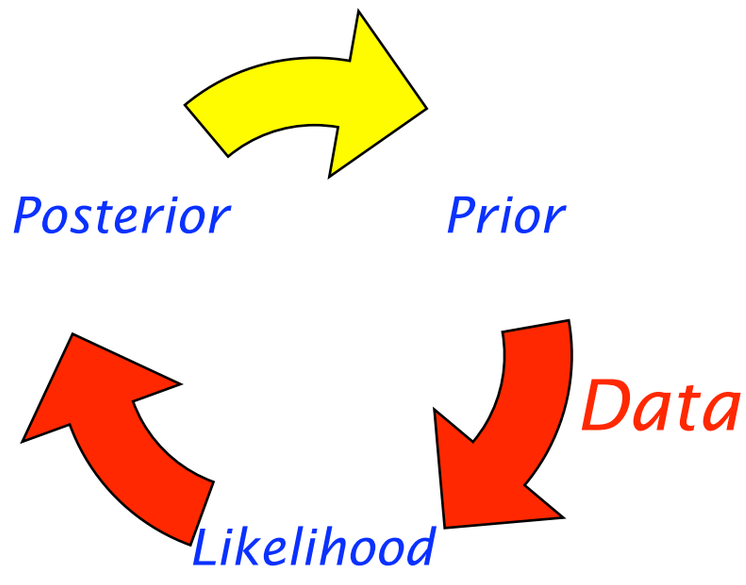
- In parameter inference, prior dependence will **in principle** vanish for strongly constraining data.
  A sensitivity analysis is mandatory for all Bayesian methods!

Likelihood (1 datum)

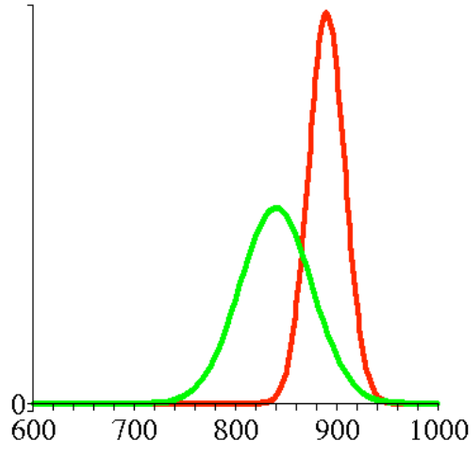


Priors

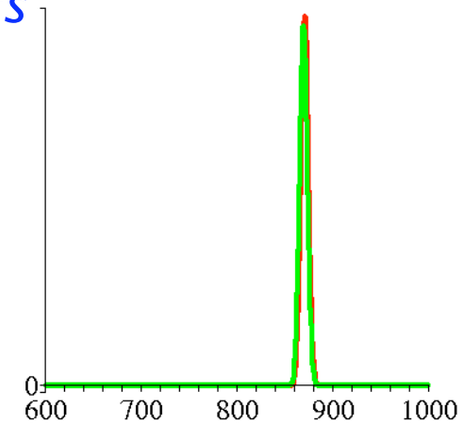




Posterior after 1 datum



Posterior after 100 data points



All the equations you'll ever need!

Imperial College London

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

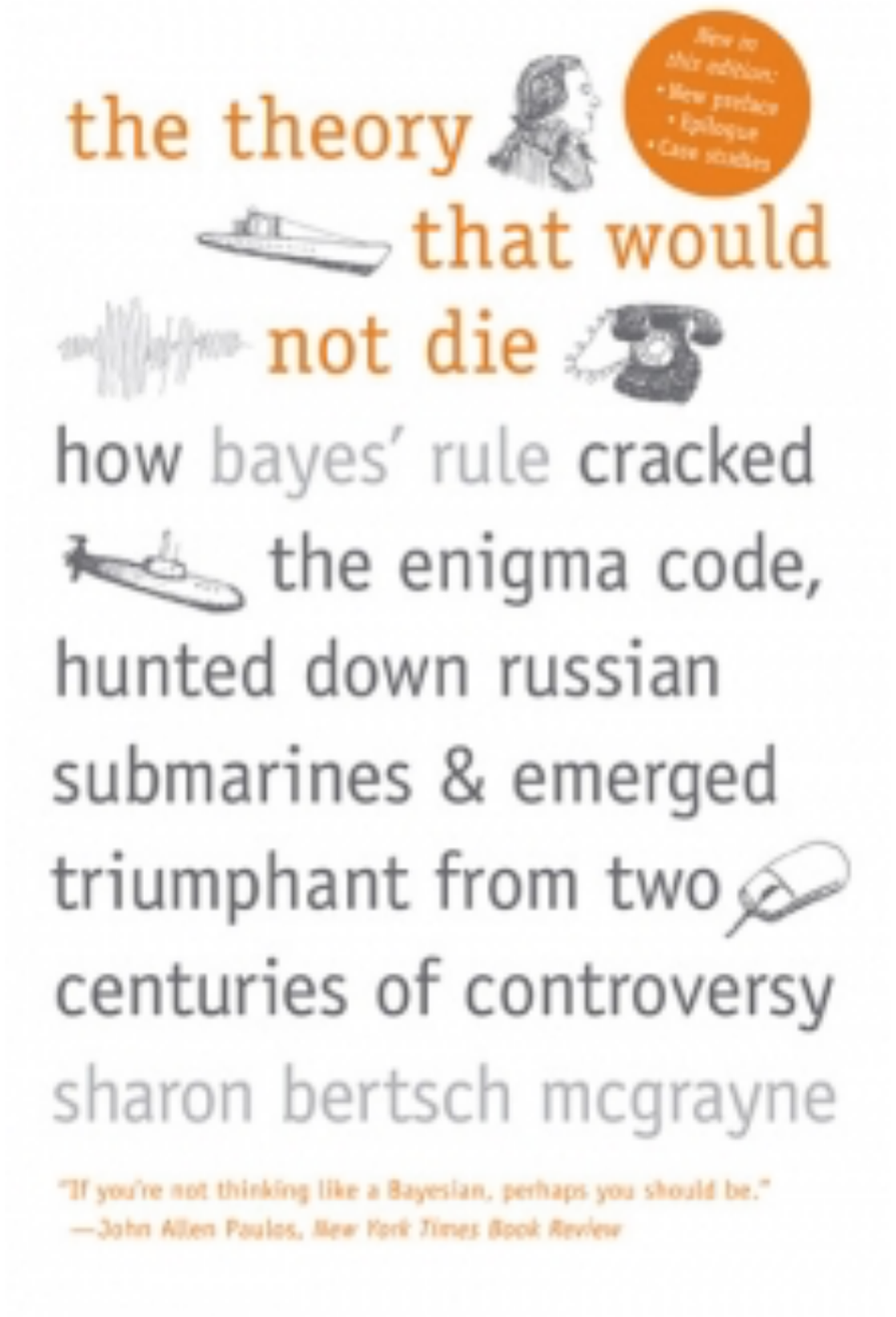
(Bayes Theorem)

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

"Expanding the discourse" or marginalisation rule

Writing the joint in terms of the conditional

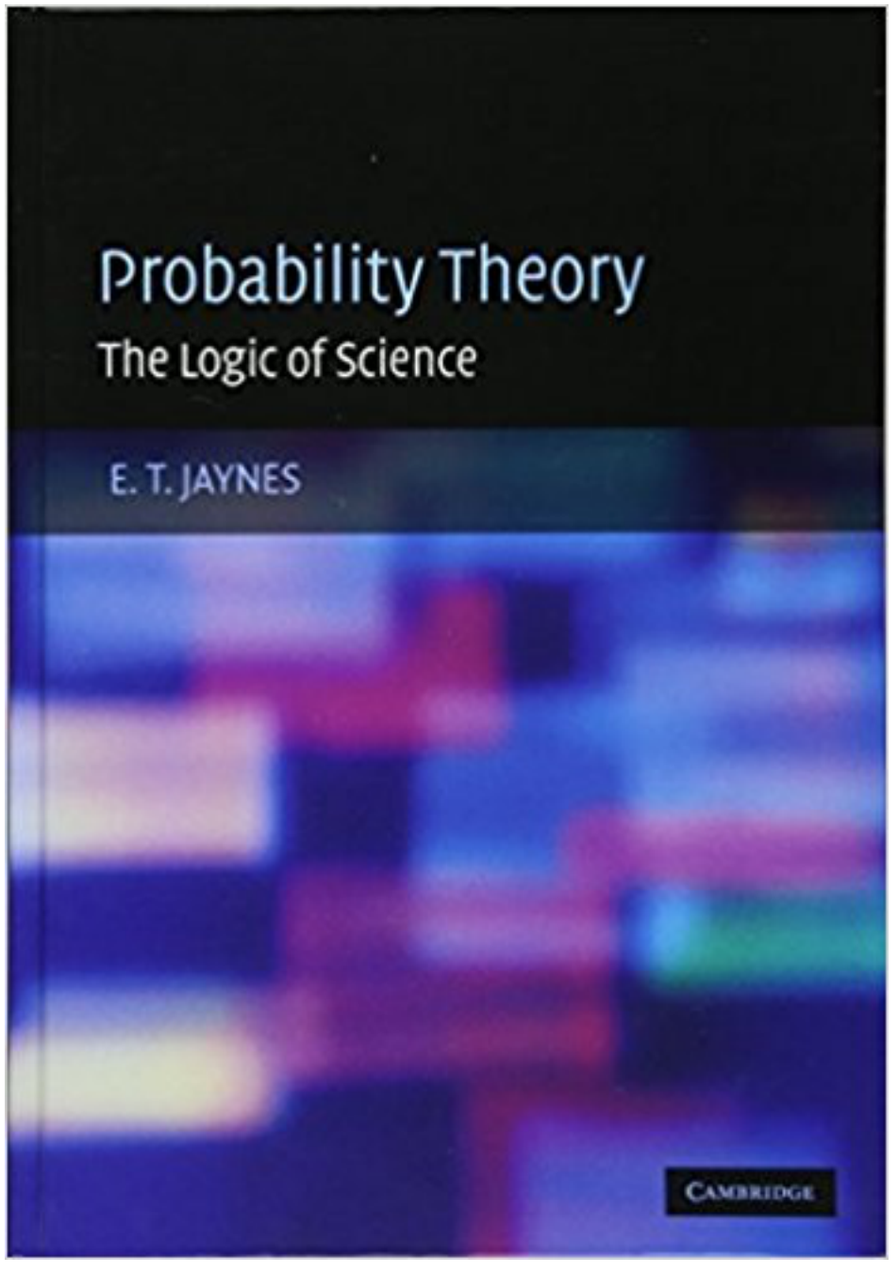
To Bayes or Not To Bayes



The Theory That Would Not Die

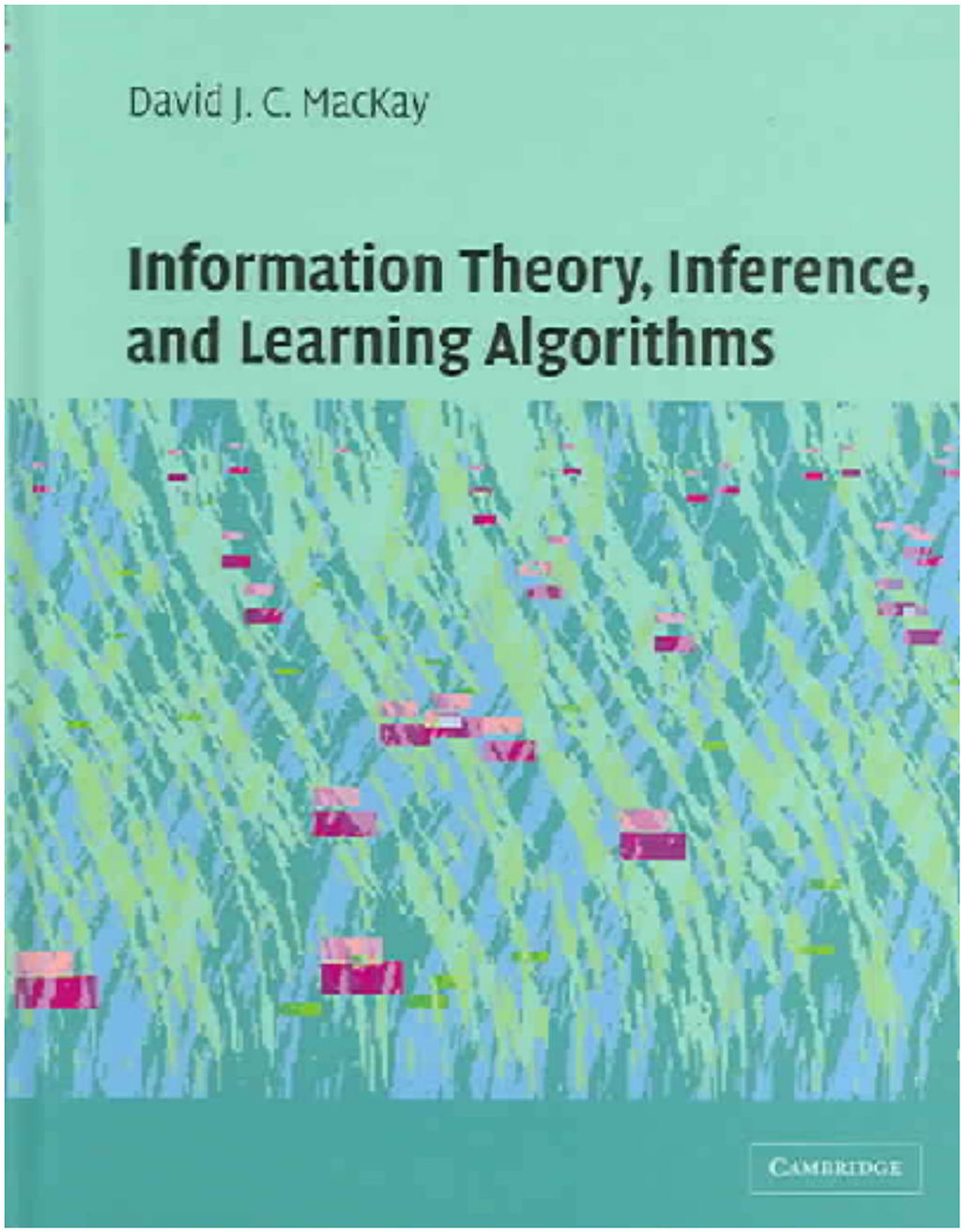
Sharon Bertsch McGrayne

How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy



Probability Theory: The Logic of Science

E.T. Jaynes



Information Theory, Inference and Learning Algorithms

David MacKay

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±0.01 mean?

Imperial College London

- **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.

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

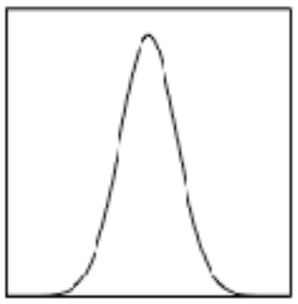
Profile likelihood:

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

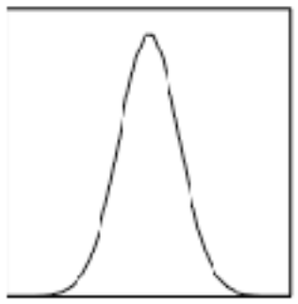
The Gaussian case

Imperial College London

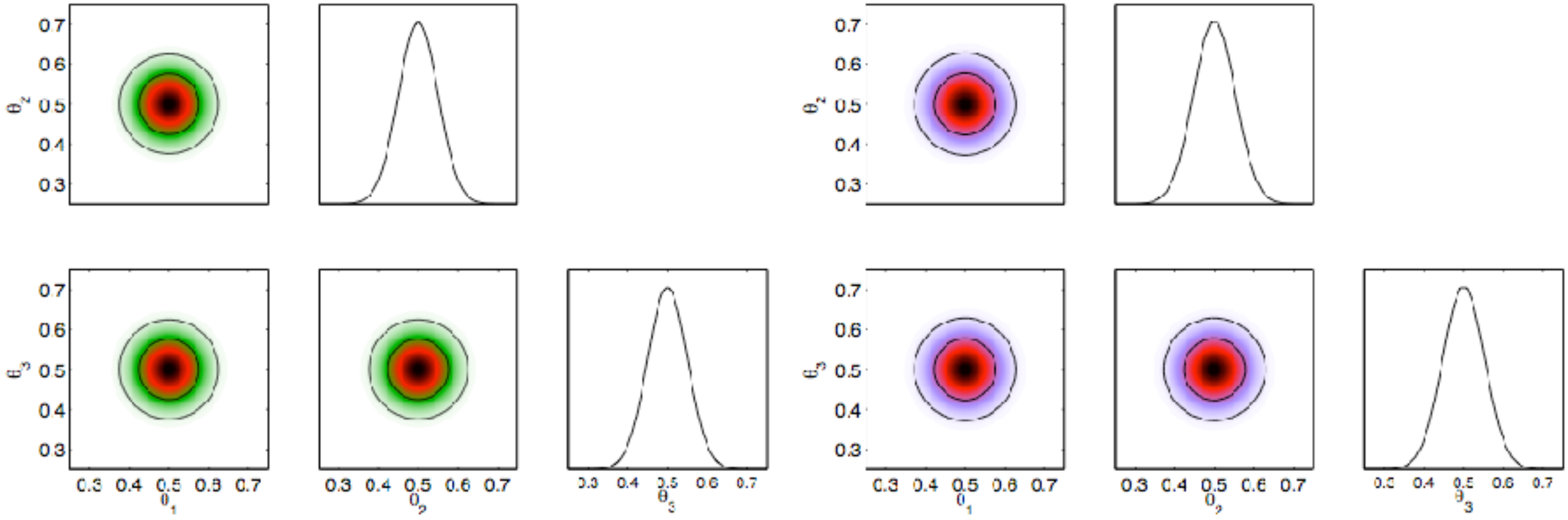
- Life is easy (and boring) in Gaussianland:



Profile likelihood



Marginal posterior



The good news

Imperial College London

- Marginalisation and profiling give exactly identical results for the linear Gaussian case.
- This is not surprising, as we already saw that the answer for the Gaussian case is numerically identical for both approaches
- And now the bad news: THIS IS NOT GENERICALLY TRUE!
- A good example is the **Neyman-Scott problem:**
  - We want to measure the signal amplitude μ_i of *N* sources with an uncalibrated instrument, whose Gaussian noise level σ is constant but unknown.
  - Ideally, measure the amplitude of calibration sources or measure one source many times, and infer the value of σ

Neyman-Scott problem

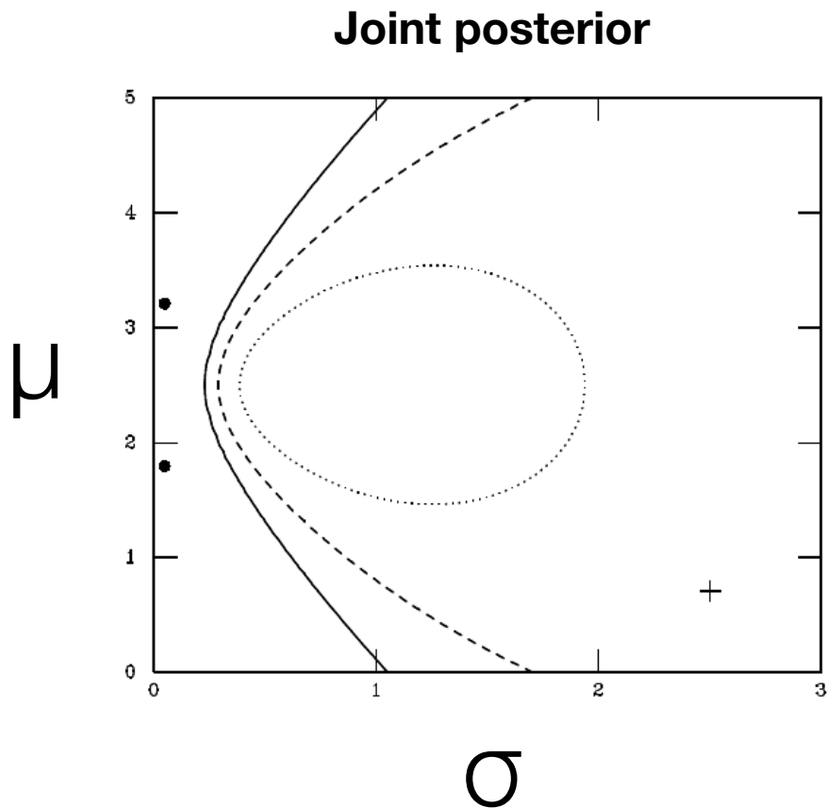
Imperial College London

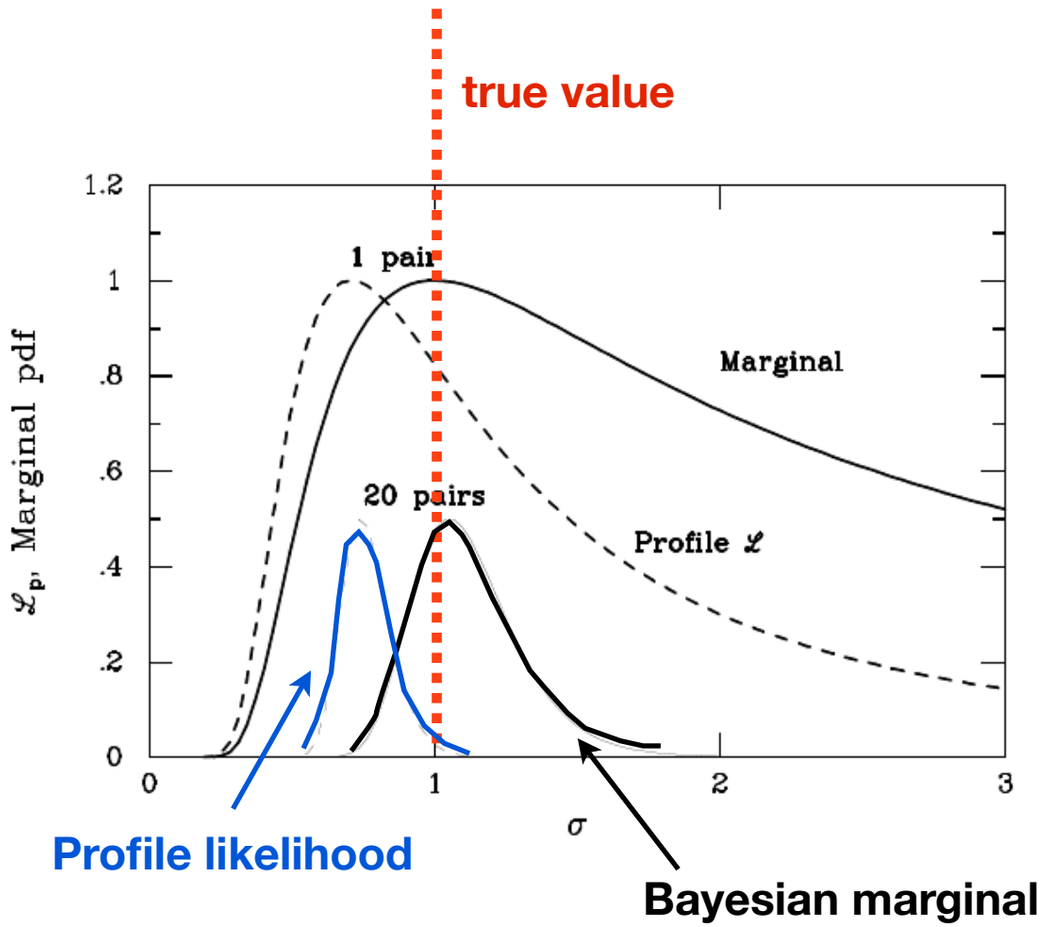
- In the Neyman-Scott problem, no calibration source is available and we can only get 2 measurements per source. So for N sources, we have N+1 parameters and 2N data points.
- The profile likelihood estimate of σ converges to a biased value σ/sqrt(2) for N → ∞
- The Bayesian answer has larger variance but is unbiased

Neyman-Scott problem

Imperial College London

Tom Loredo, talk at Banff 2010 workshop:

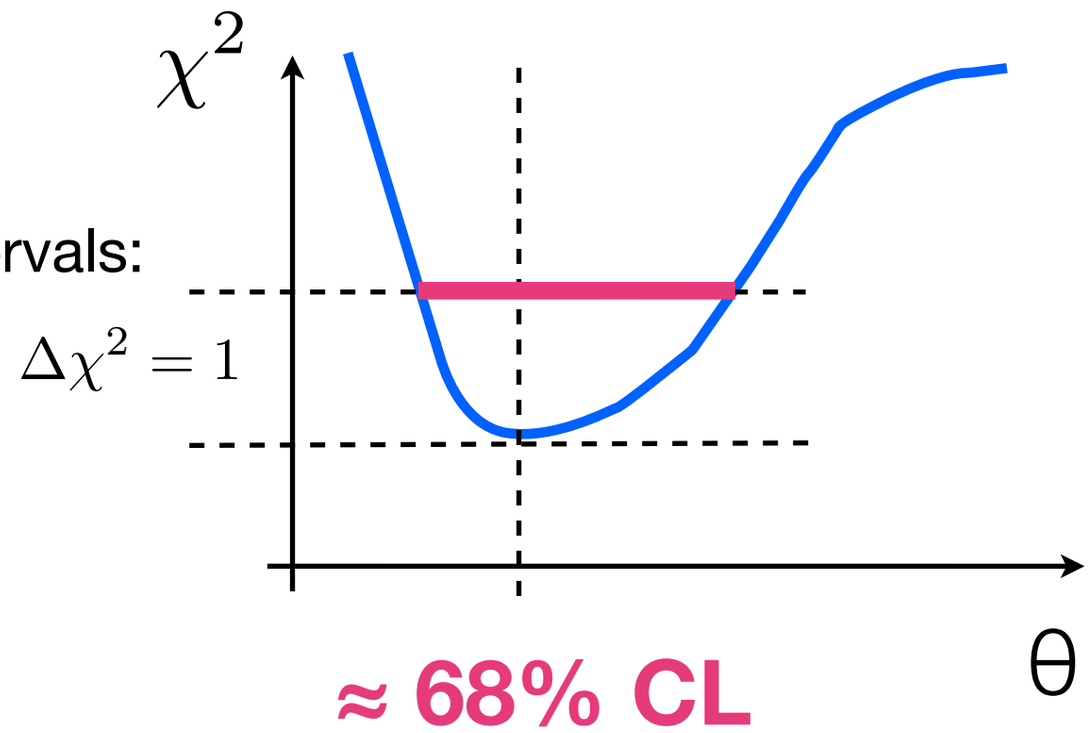




Confidence intervals: Frequentist approach

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

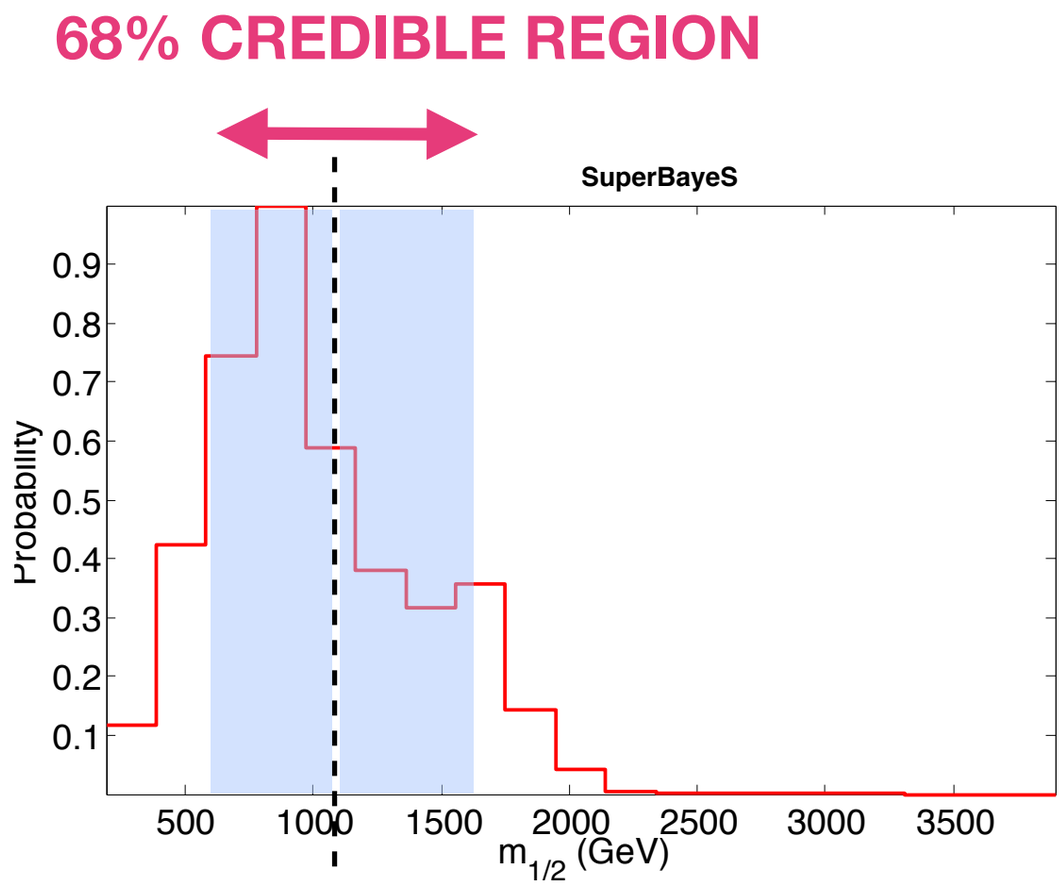


≈ 68% CL

Credible regions: Bayesian approach

Imperial College London

- 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

Imperial College London

- 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)

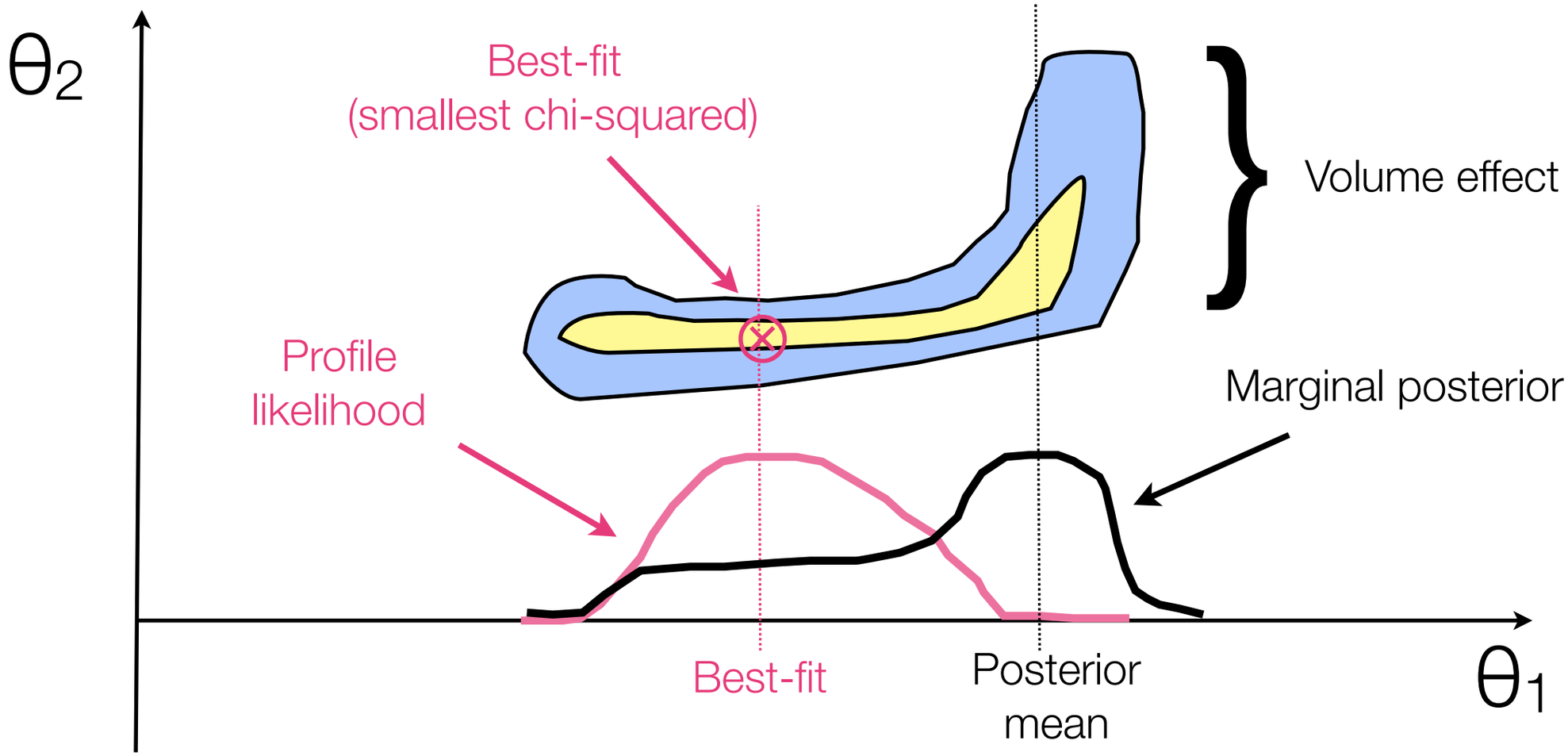
Imperial College London

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)

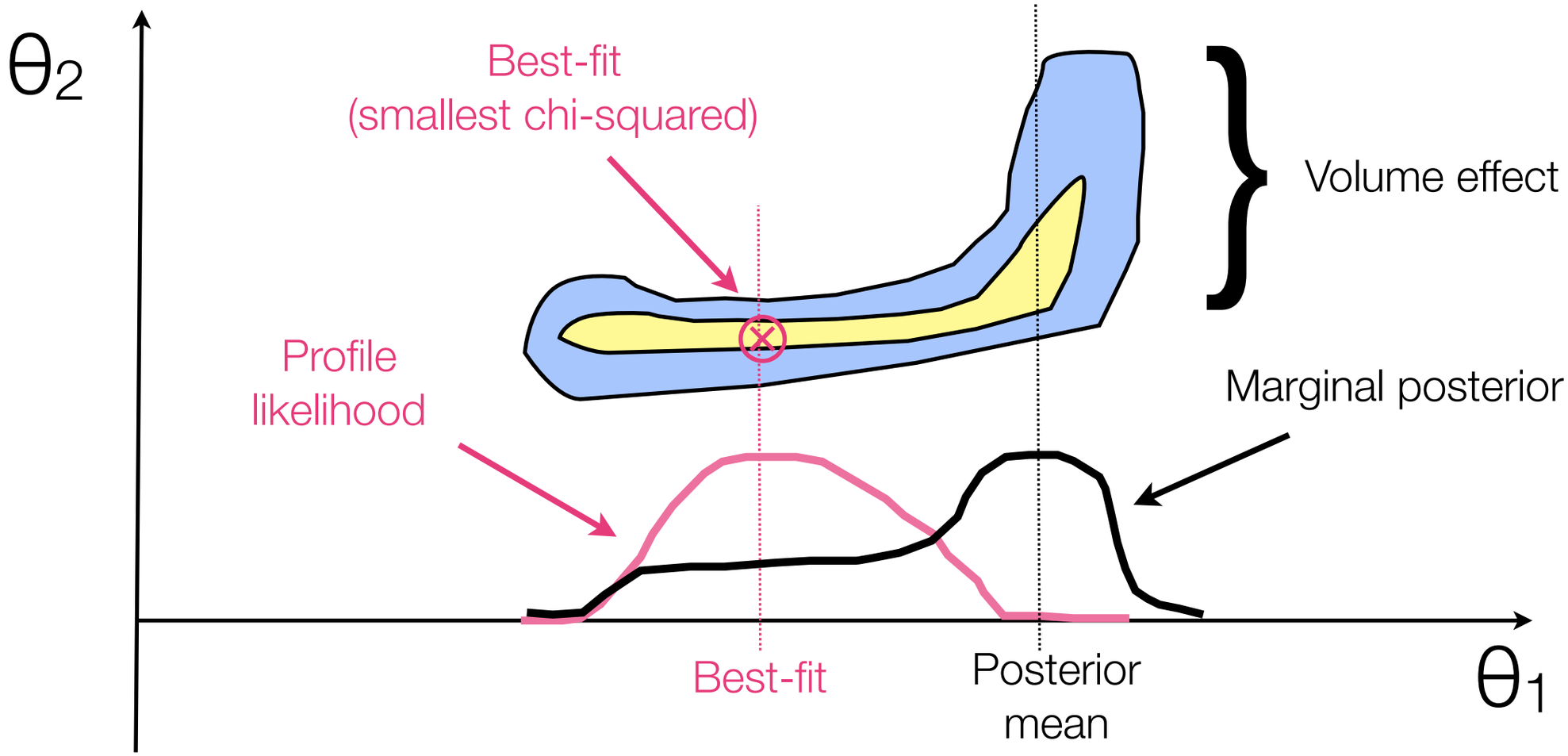
Imperial College London

Physical analogy: (thanks to Tom Loredo)

Likelihood = hottest hypothesis
Posterior = hypothesis with most heat

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

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



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

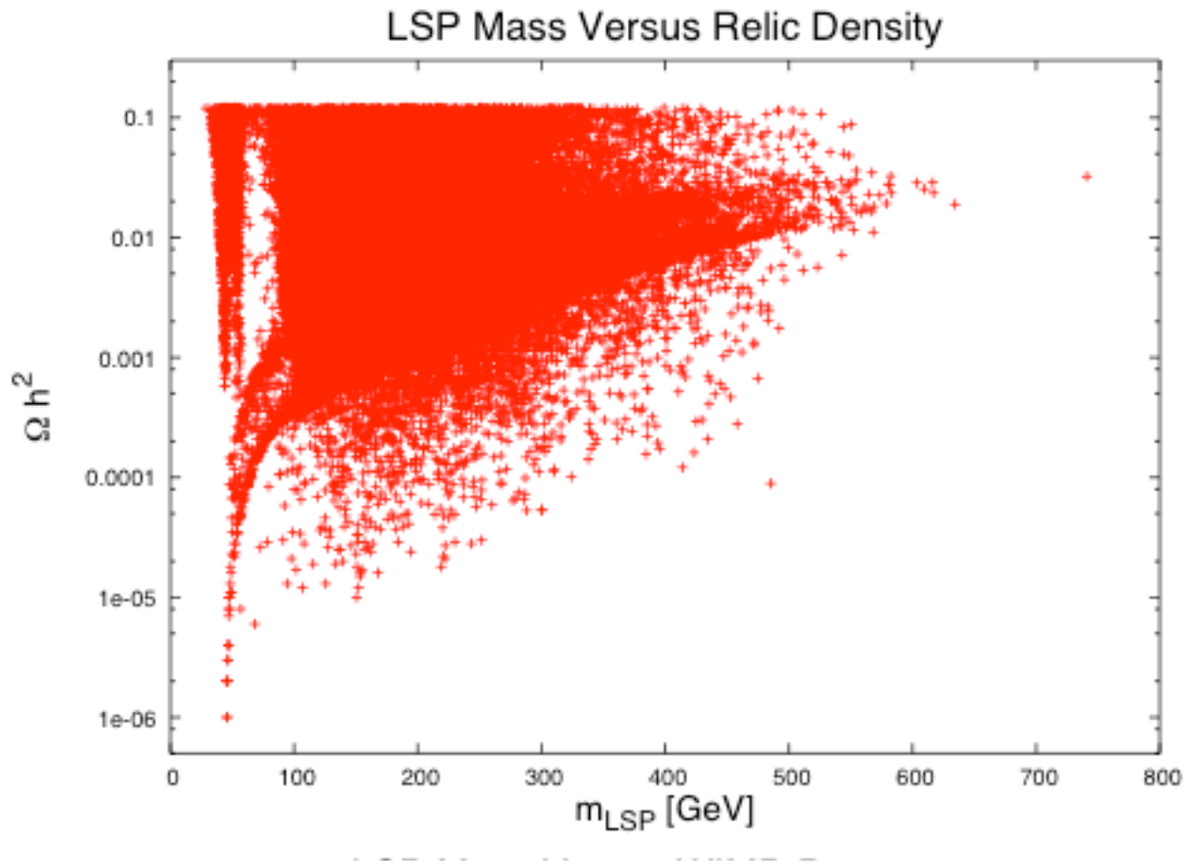
Markov Chain Monte Carlo

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!

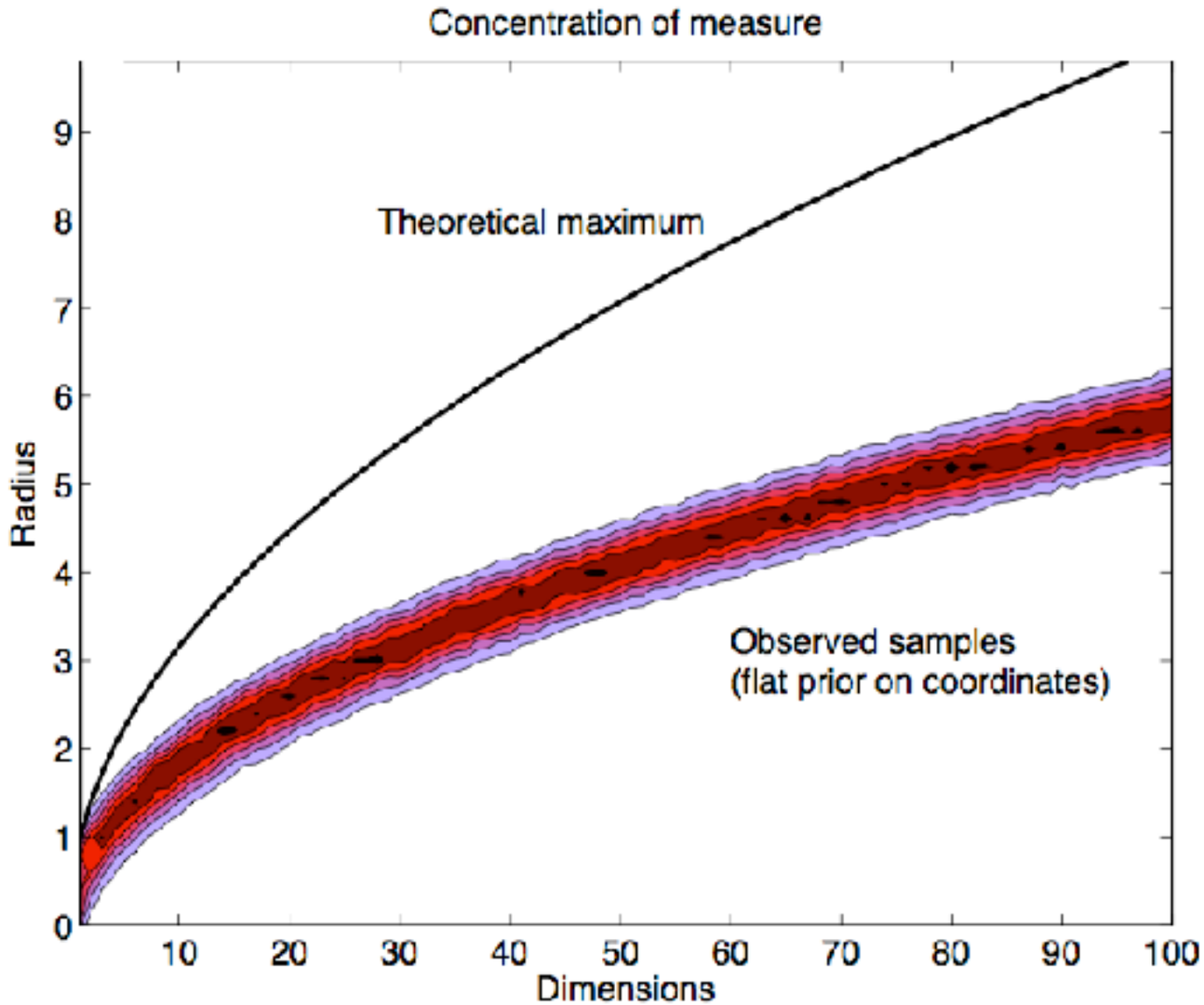
One example:
Berger et al (0812.0980)
pMSSM scans
(20 dimensions)

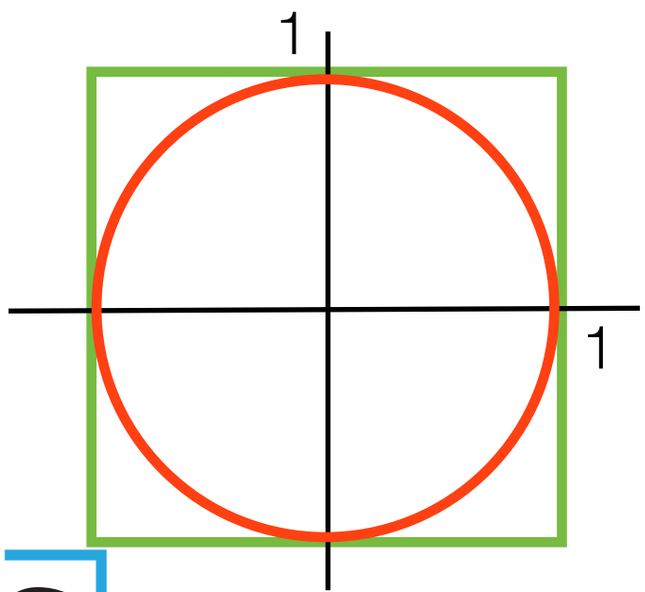


Random scans explore only a small fraction of the parameter space

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





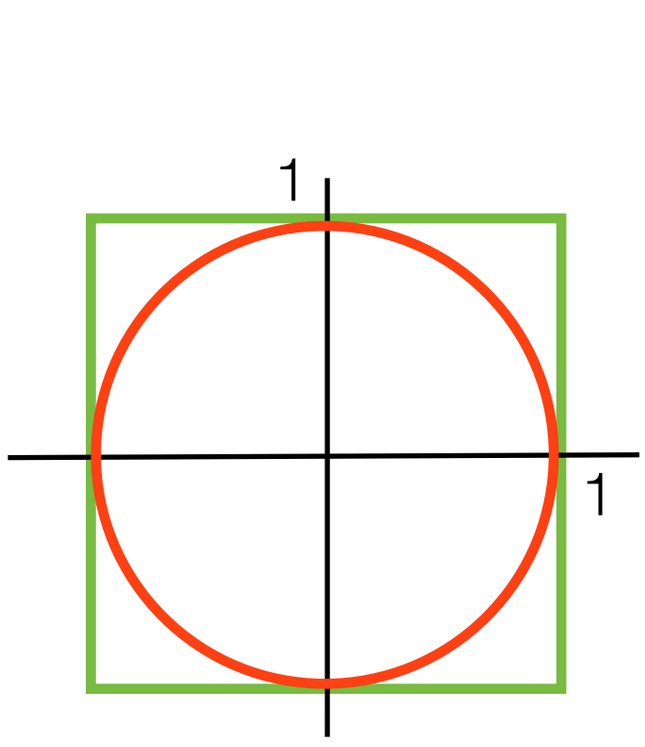
ICIC

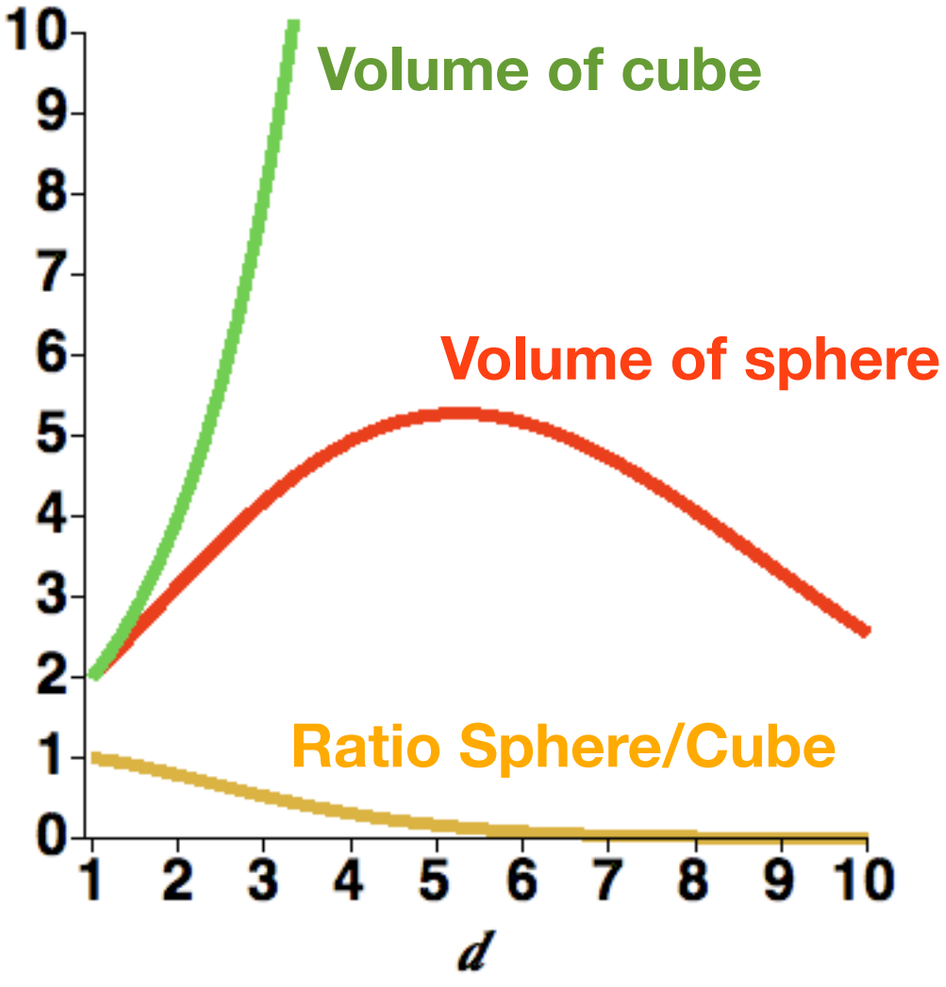
Geometry in high-D spaces

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.





ICIC

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

Imperial College London

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

- 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

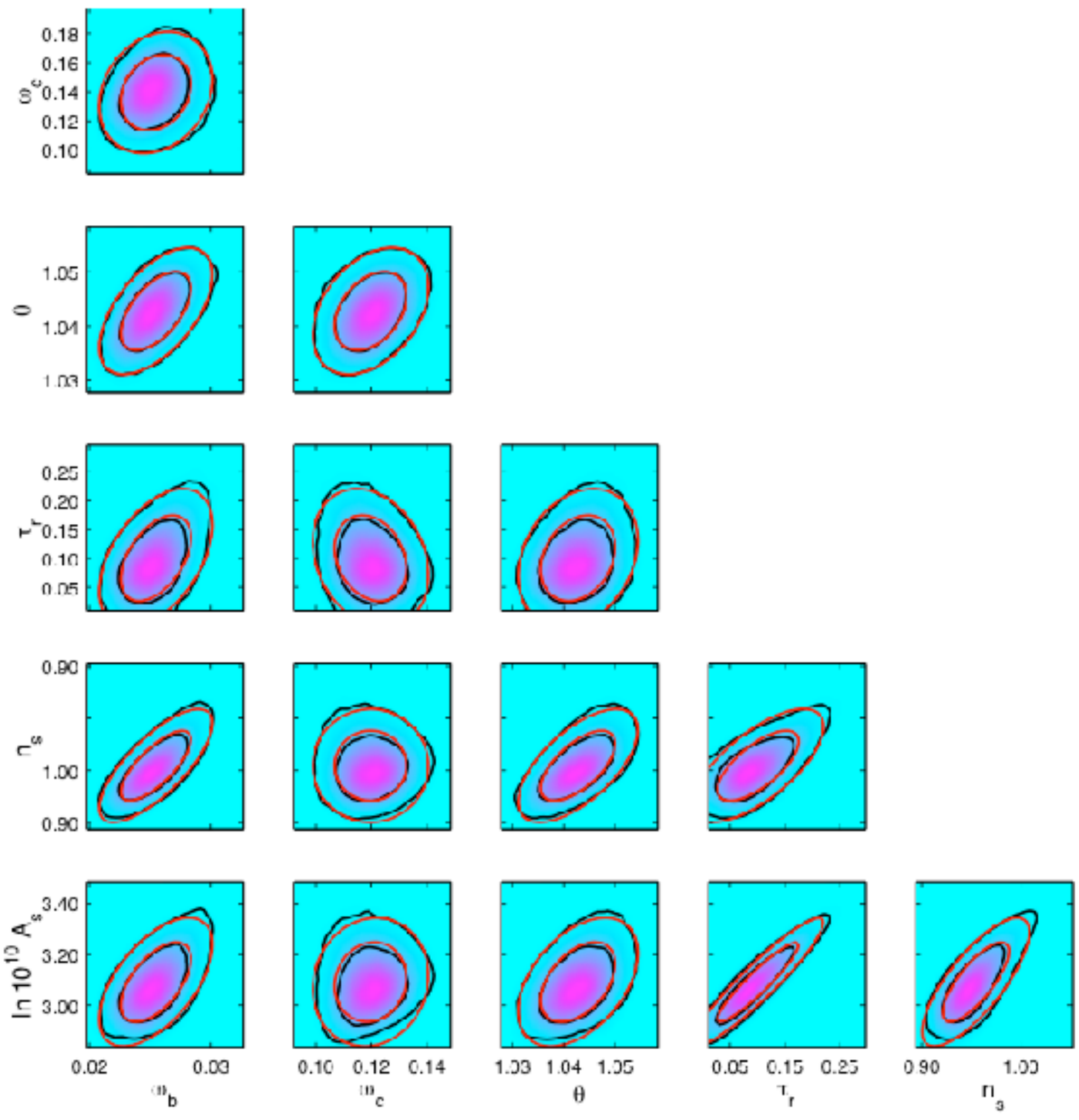
Imperial College London

- **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

Imperial College London

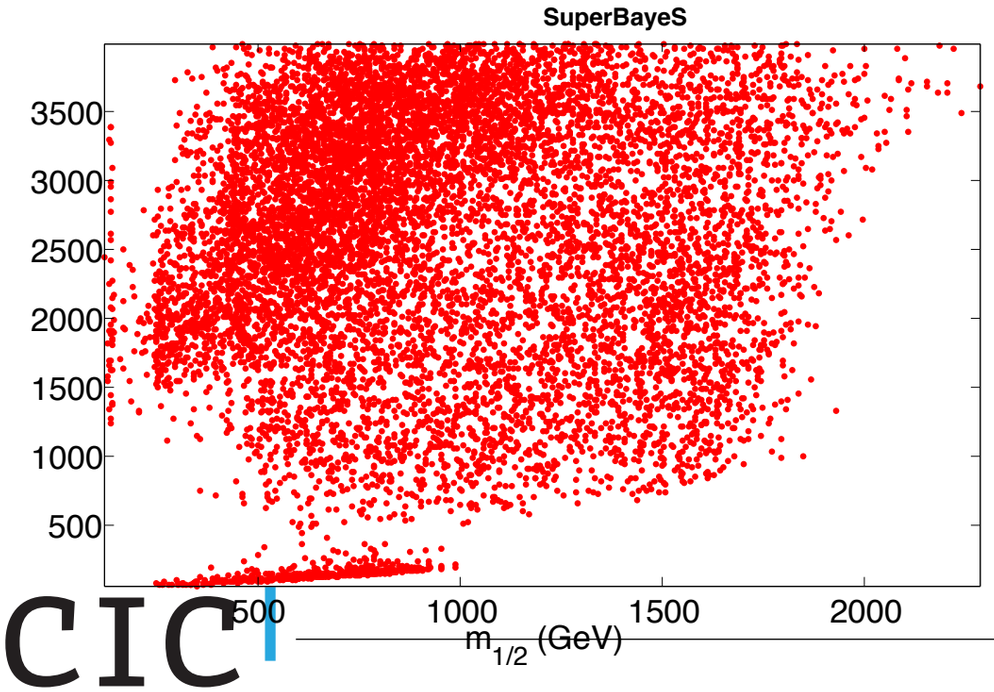


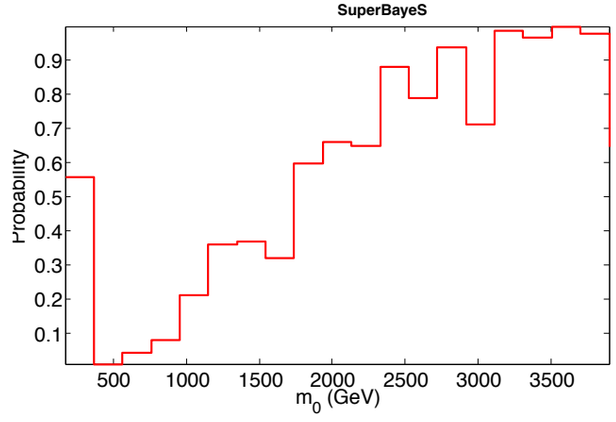
ICIC

MCMC estimation

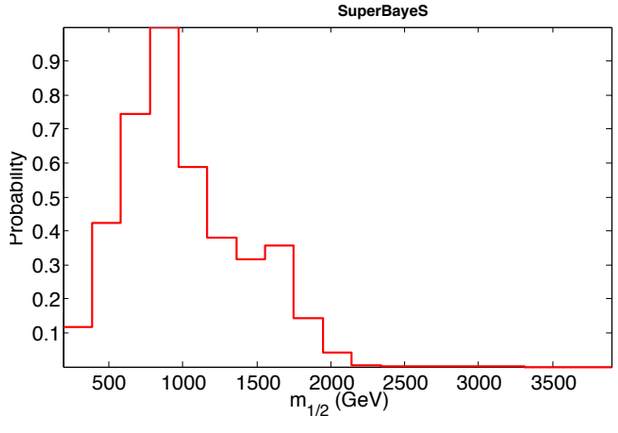
- **Marginalisation becomes trivial:** create bins along the dimension of interest and simply count samples falling within each bins ignoring all other coordinates
- Examples (from **superbayes.org**) :

2D distribution of samples from joint posterior





1D marginalised posterior (along y)

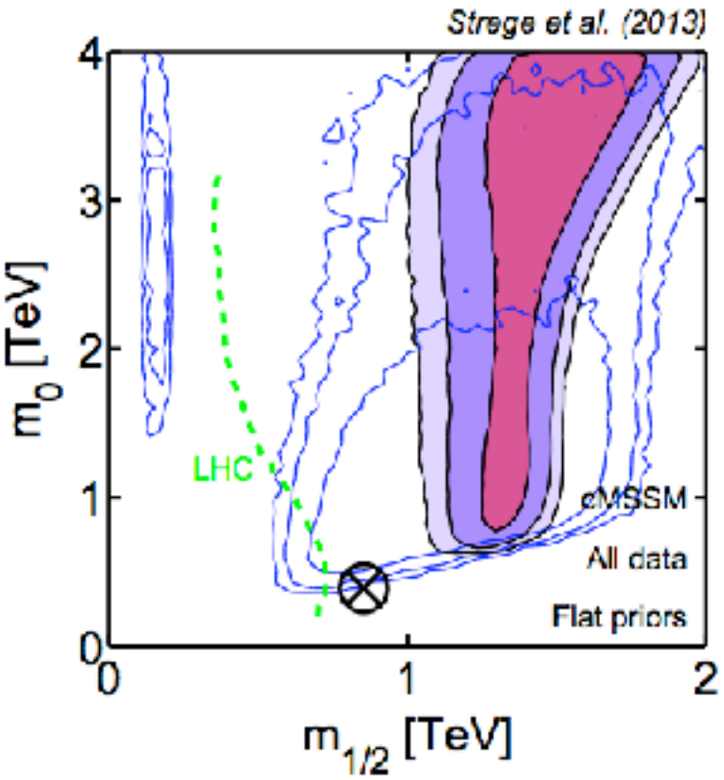


1D marginalised posterior (along x)

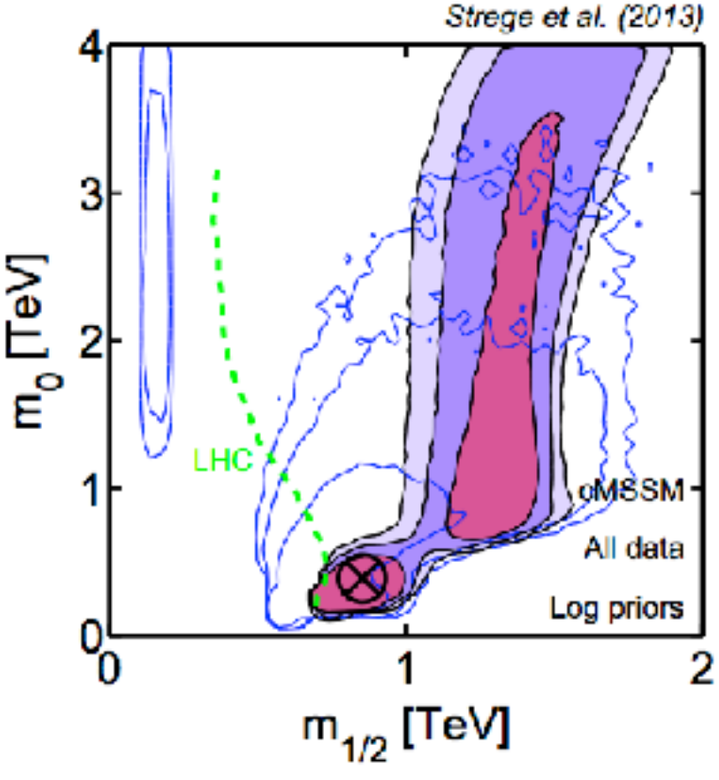
Non-Gaussian example

Imperial College London

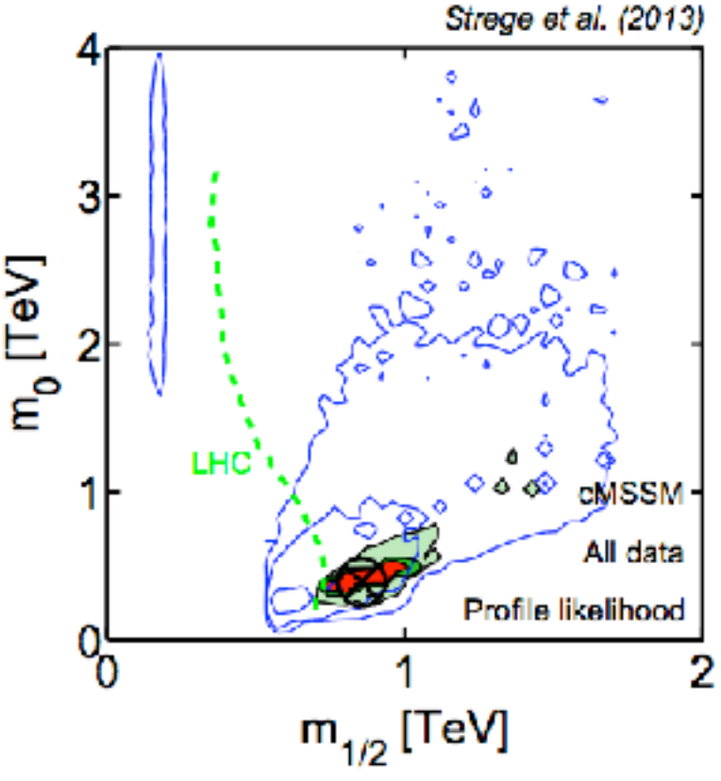
Bayesian posterior
("flat priors")



Bayesian posterior
("log priors")



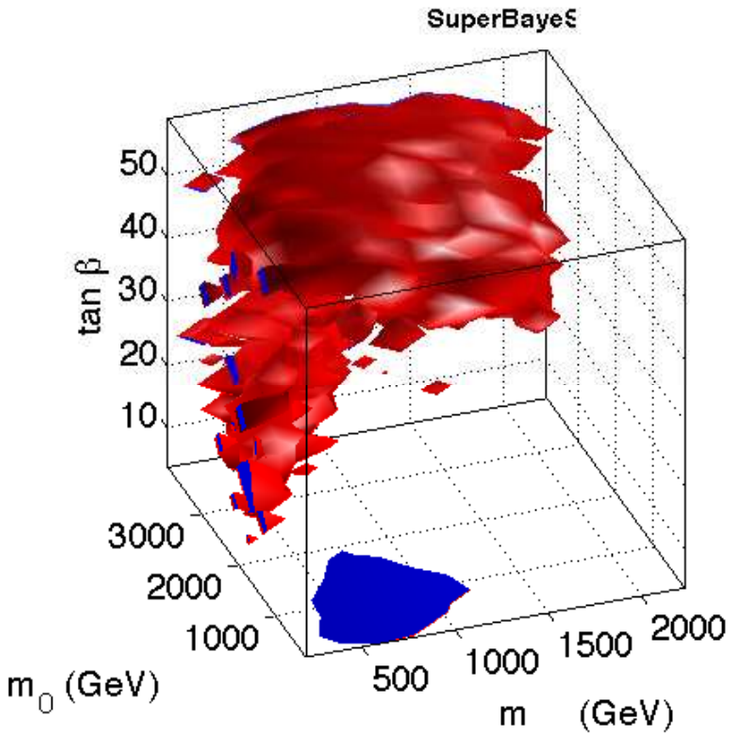
Profile likelihood

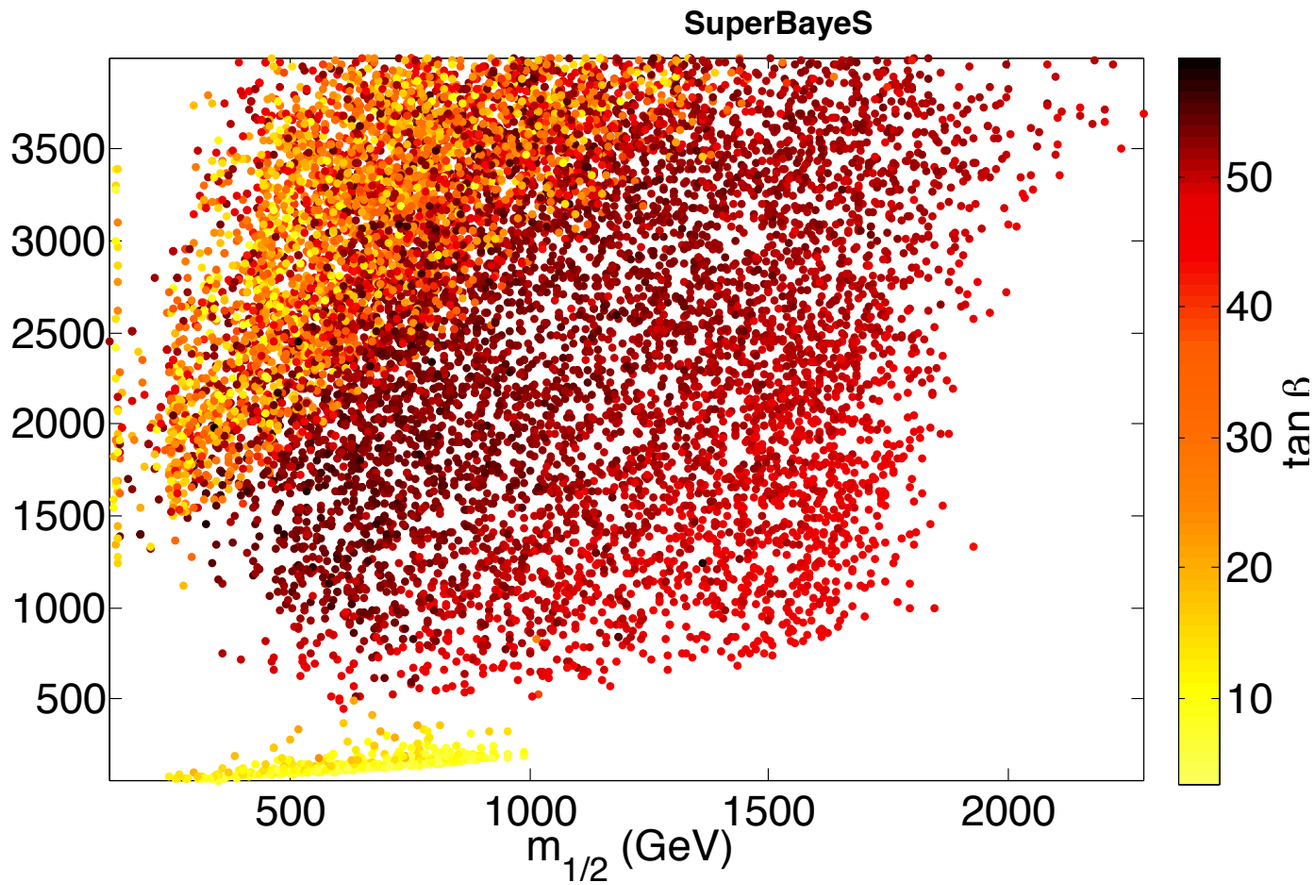


Constrained Minimal Supersymmetric Standard Model (4 parameters)
Strege, RT et al (2013)

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

Imperial College London

- 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

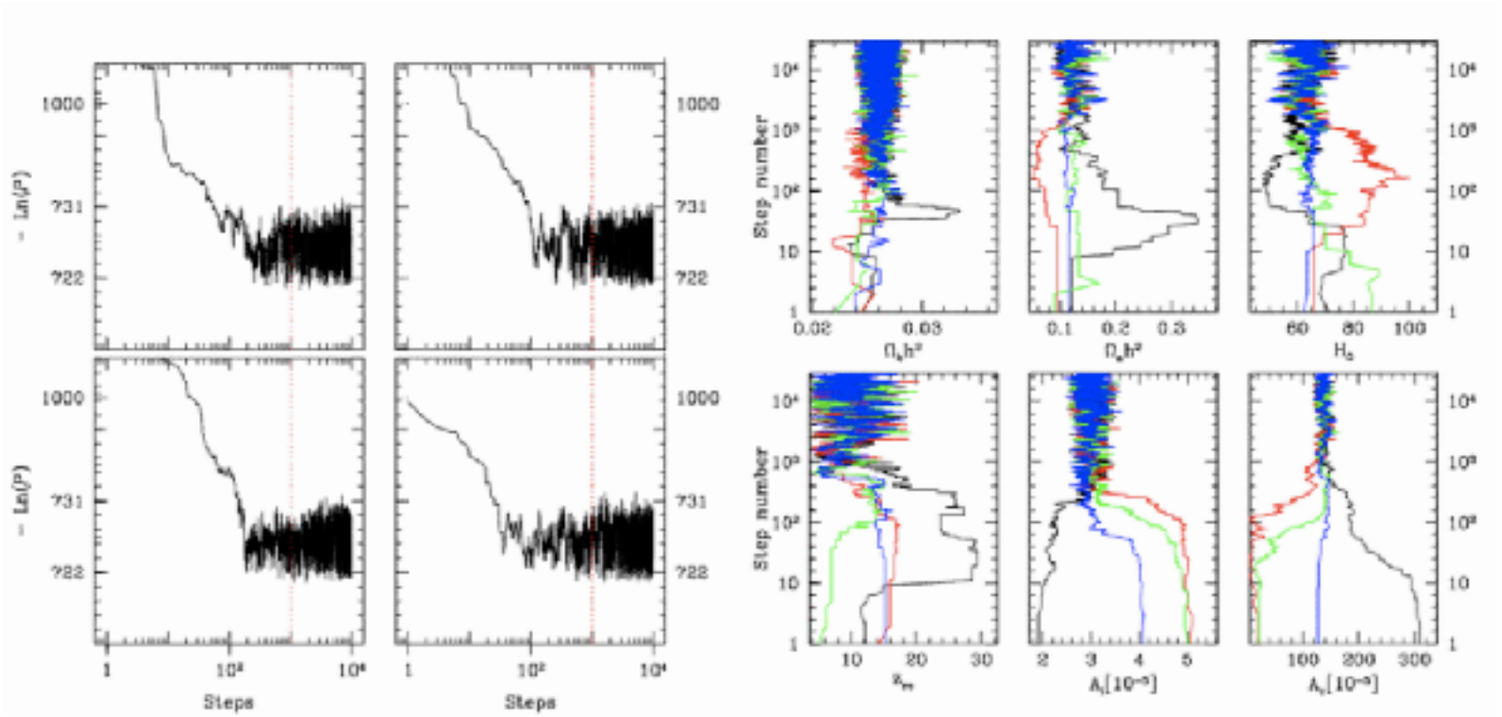
MCMC diagnostics

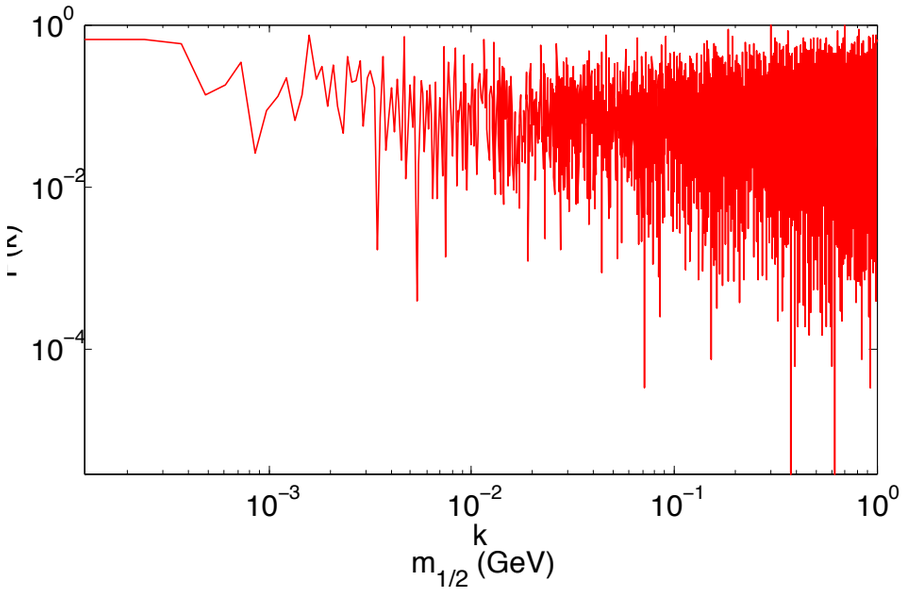
Imperial College London

Burn in

Mixing

Power spectrum





(see astro-ph/0405462 for details)

ICIC

MCMC samplers you might use

Imperial College London

- **PyMC** Python package: https://pymc-devs.github.io/pymc/
  Implements Metropolis-Hastings (adaptive) MCMC; Slice sampling; Gibbs sampling.
  Also has methods for plotting and analysing resulting chains.

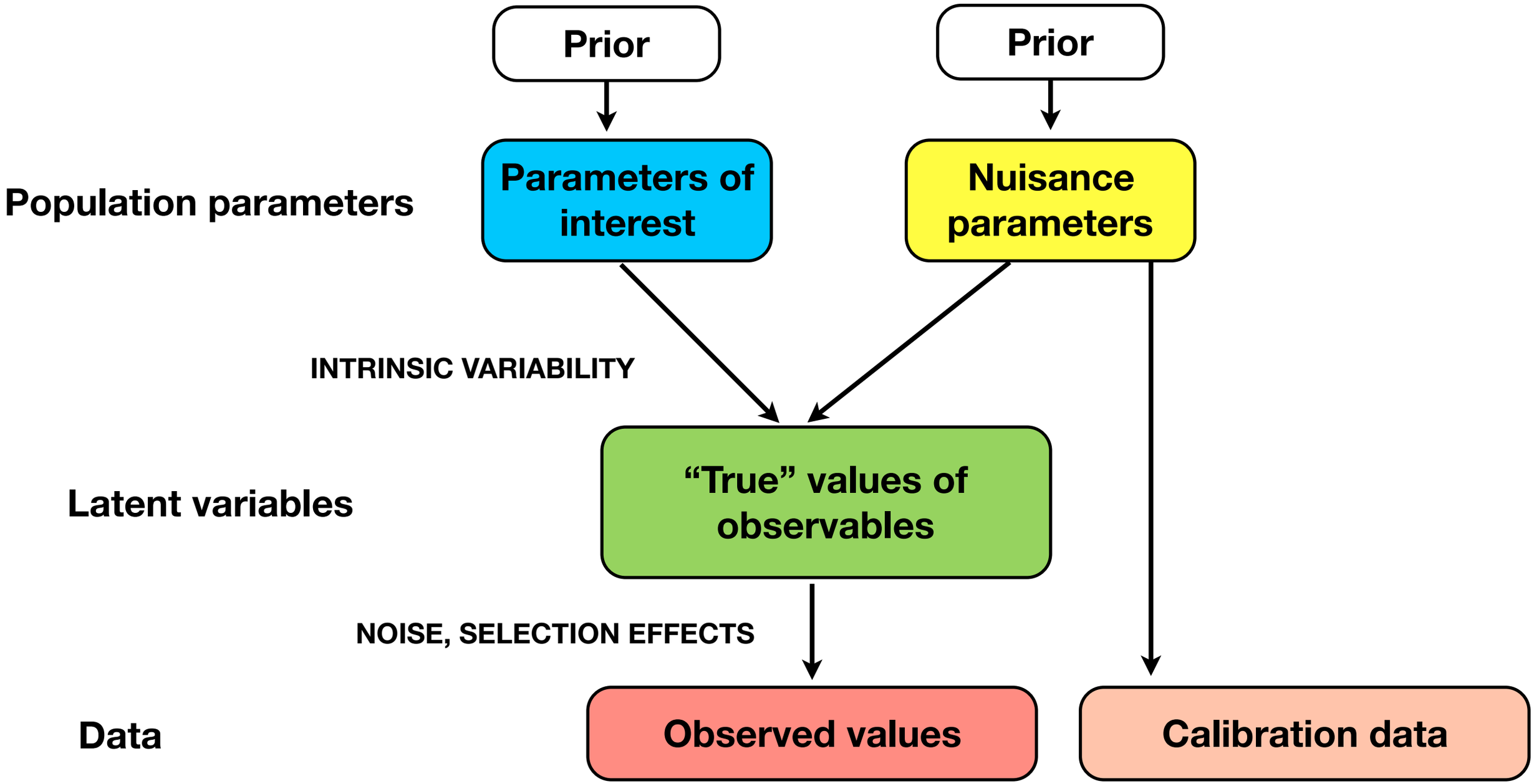
- **emcee** ("The MCMC Hammer"): http://dan.iel.fm/emcee
  Dan Foreman-Makey et al. Uses affine invariant MCMC ensemble sampler.

- **Stan** (includes among others Python interface, PyStan): http://mc-stan.org/
  Andrew Gelman et al. Uses Hamiltonian MC.

- Practical example of straight line regression, installation tips and comparison between the 3 packages by Jake Vanderplas: http://jakevdp.github.io/blog/2014/06/14/frequentism-and-bayesianism-4-bayesian-in-python/
  (check out his blog, *Pythonic Preambulations*)

Bayesian hierarchical models

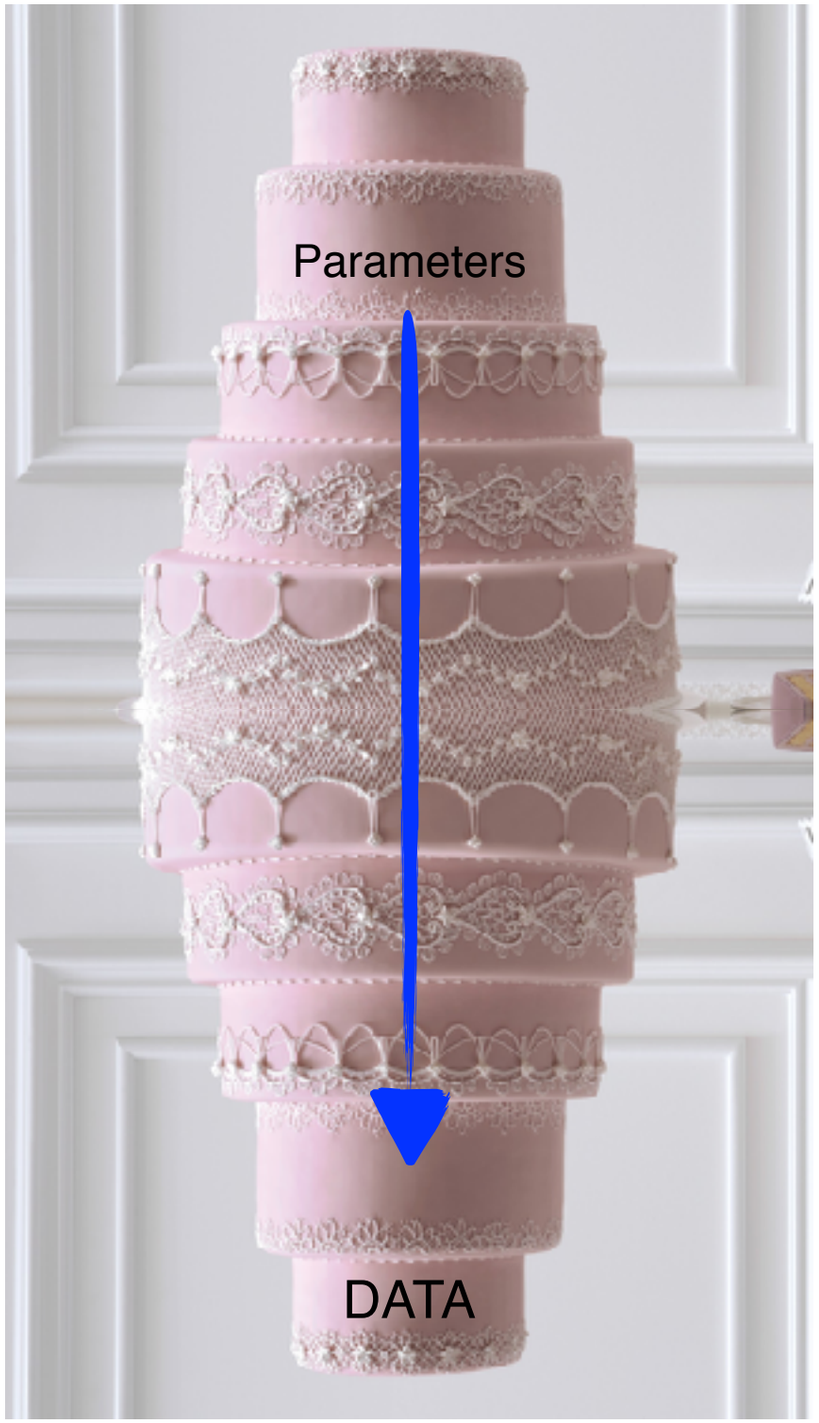
Imperial College London



Why “Hierarchical”?

Imperial College London

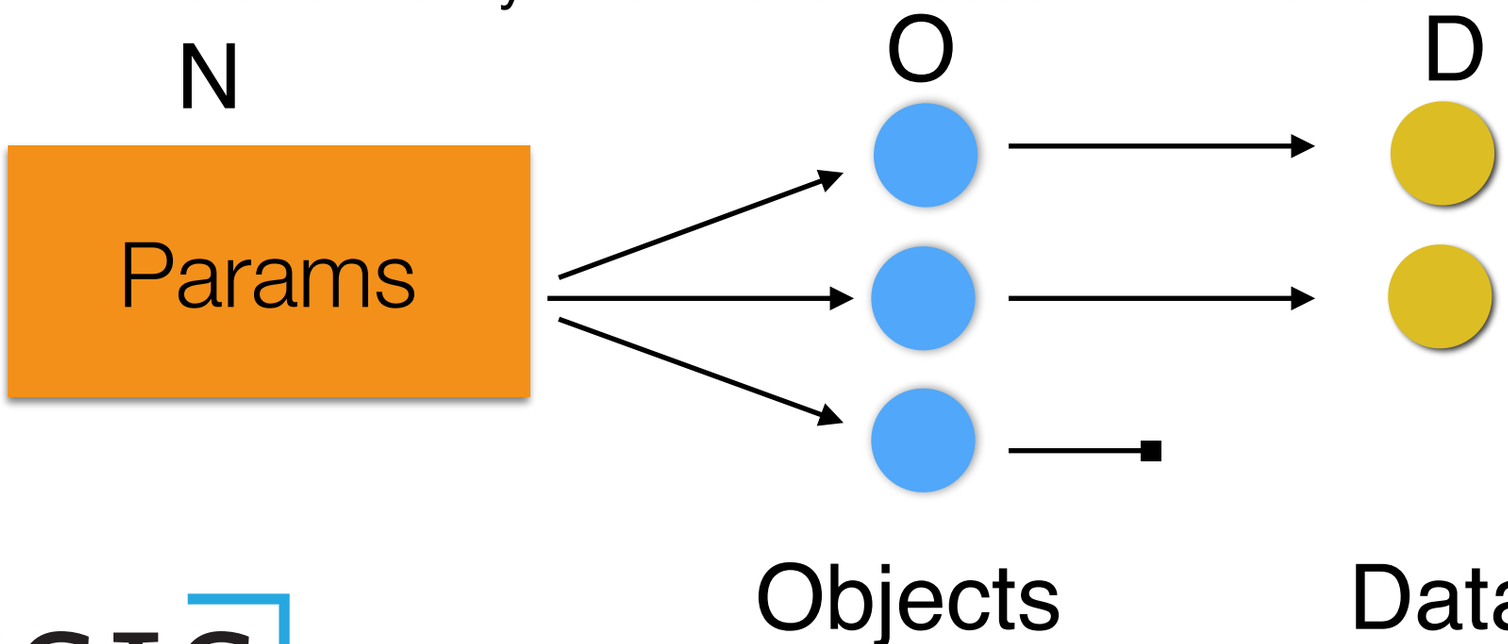
- In cosmology, we have many problems of interest where the “objects” of study are used as tracers for underlying phenomena
- Eg:
  - SNIa’s to measure d_L
  - Galaxies to measure velocity fields, BAOs, growth of structure, lensing, …
  - Galaxy properties to measure scaling relationships
  - Stars to measure Milky Way gravitational potential/dark matter
  - …
- In many cases, we might or might not be interested in the objects themselves — insofar as they give us accurate (and unbiased) tracers for the physics we want to study



Why “Models”?

Imperial College London

- By “model” in this context I mean a probabilistic representation of how the measured data arise from the theory
- We always need models: They incorporate our understanding of how the measurement process (and its subtleties, e.g. section effects) “filters” our view of the underlying physical process
- The more refined the model, the more information we can extract from the data: measurement noise is unavoidable (at some level), but supplementing our inferential setup with a probabilistic model takes some “heavy lifting” away from the data
- The key is to realise that there is a difference between “measurement noise” and intrinsic variability — and each needs to be modelled individually



Mathematical formulation

Imperial College London

The posterior distribution can be expanded in the usual Bayesian way:

$$p(\text{params} \mid \text{data}) \propto p(\text{data} \mid \text{params})p(\text{params})$$

$$p(\text{data} \mid \text{params}) \propto \int p(\text{data}, \text{true}, \text{pop} \mid \text{params})\, d\text{true}\, d\text{pop}$$

$$= \int p(\text{data} \mid \text{true})\, p(\text{true} \mid \text{pop})\, p(\text{pop})\, d\text{true}\, d\text{rop}$$

Measurement errors

Intrinsic variability

Population-level priors

ICIC

Gaussian linear model

Imperial College London

- Intuition can be gained from the “simple” 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):

$$y_i = b + a x_i$$

Model: unknown parameters of interest (a,b)

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

POPULATION DISTRIBUTION

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

INTRINSIC VARIABILITY

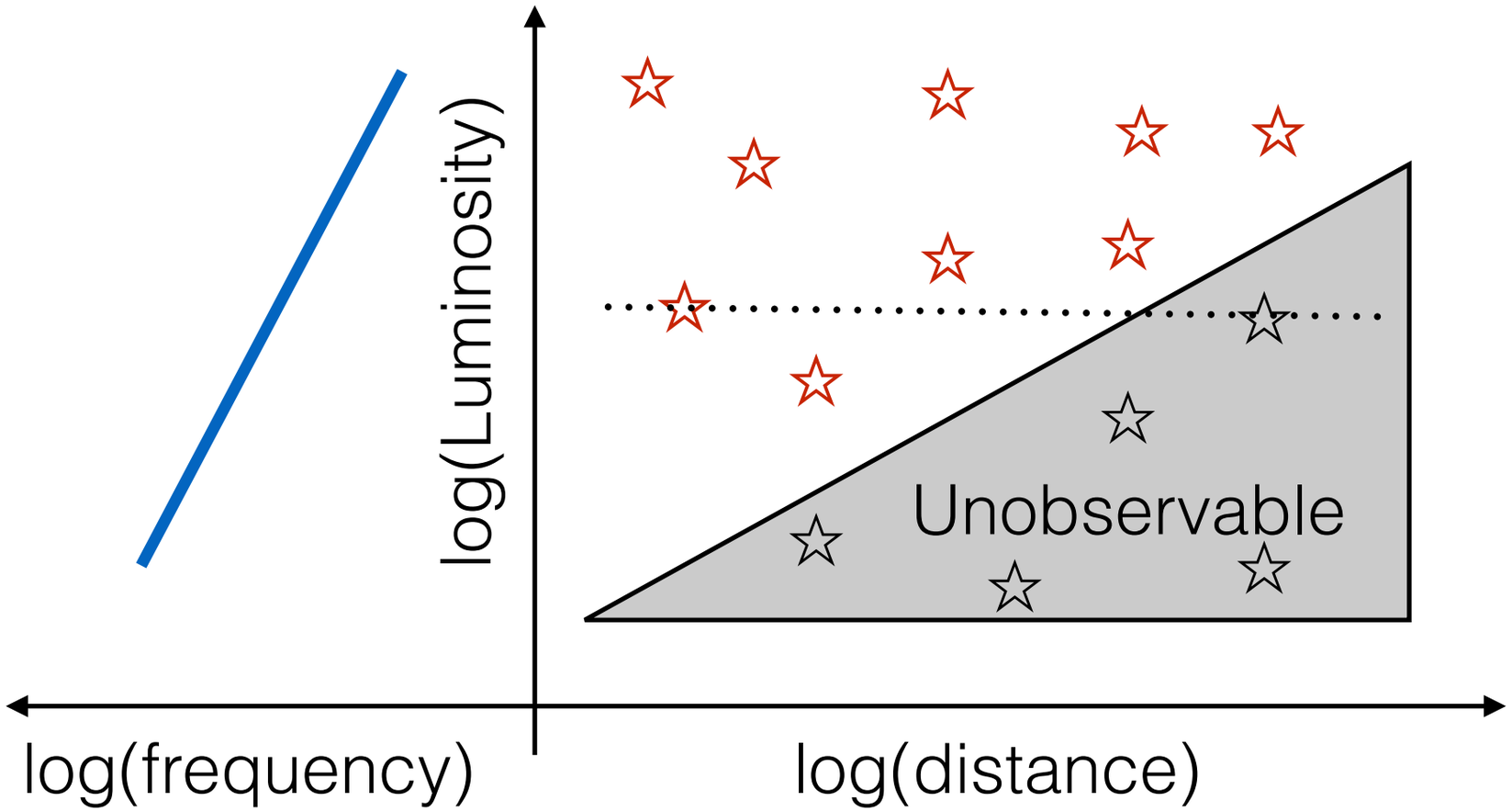
$$\hat{x}_i, \hat{y}_i|x_i, y_i \sim \mathcal{N}_{\hat{x}_i, \hat{y}_i}([x_i, y_i], \Sigma^2)$$

MEASUREMENT ERROR

Malmquist bias revisited

Imperial College London

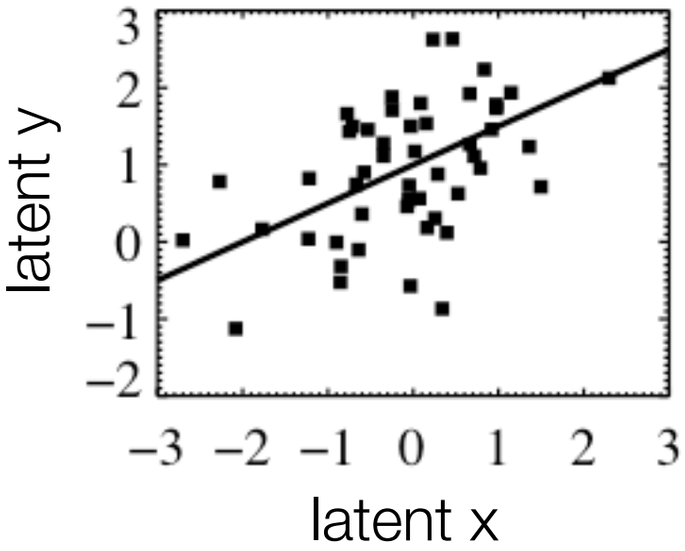
- Malmquist (1925) bias: intrinsically brighter objects are easier to detect, hence quantities derived from a magnitude (brightness) limited sample are biased high.



1. Observed objects have mean luminosity biased high

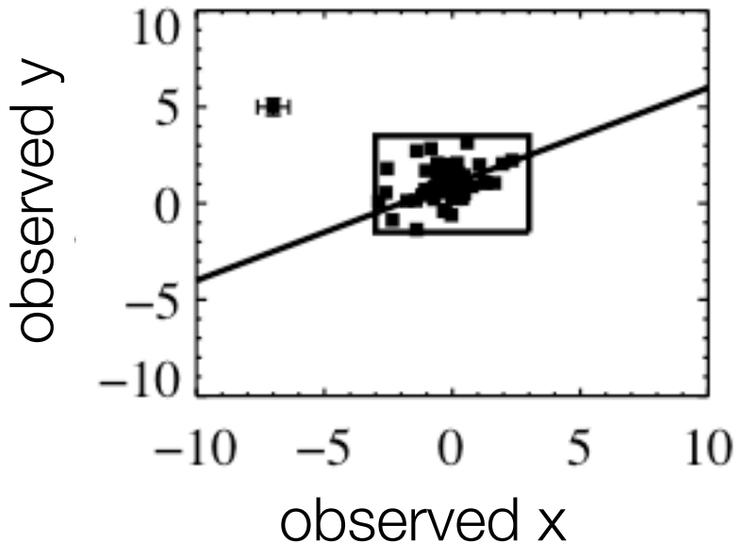
2. Noise more likely to up-scatter lower luminosity object into detection threshold than vice-versa (as less luminous objects are more frequent)

INTRINSIC VARIABILITY

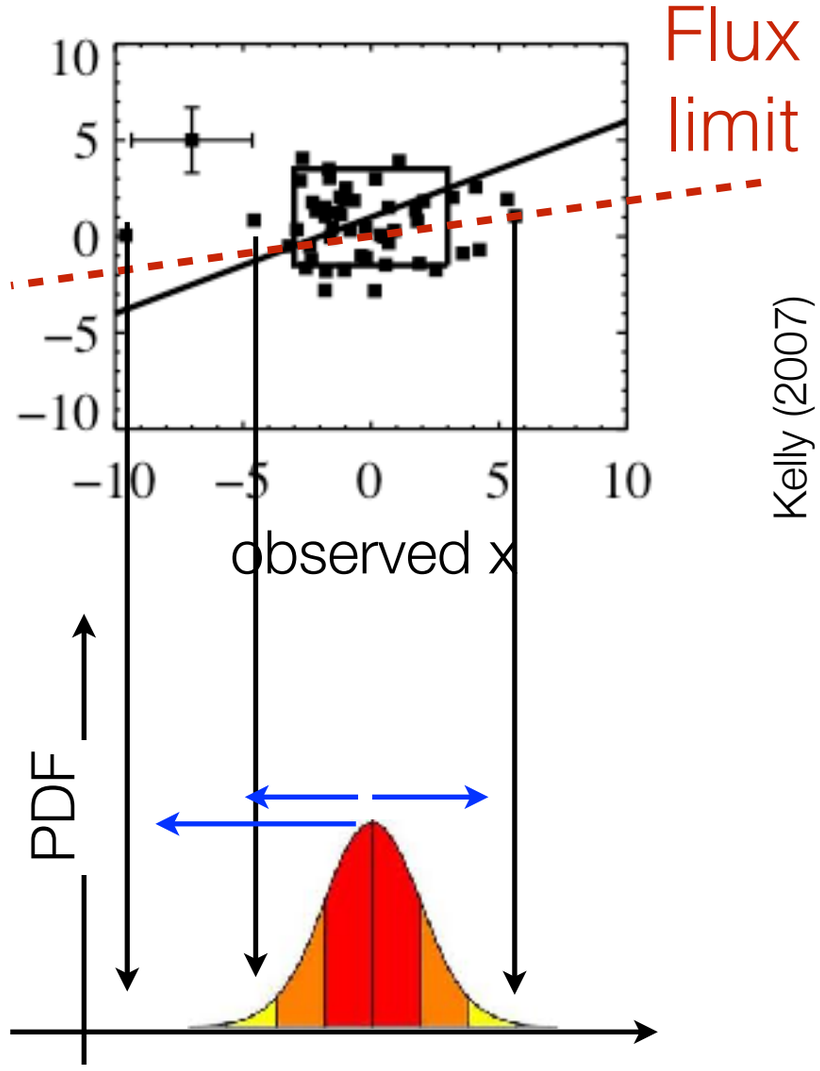


TRUE VALUES

+ MEASUREMENT ERROR



“SMALL” ERRORS



“LARGE” ERRORS

- Modeling the latent distribution of the independent variable accounts for “Malmquist bias” of the second kind

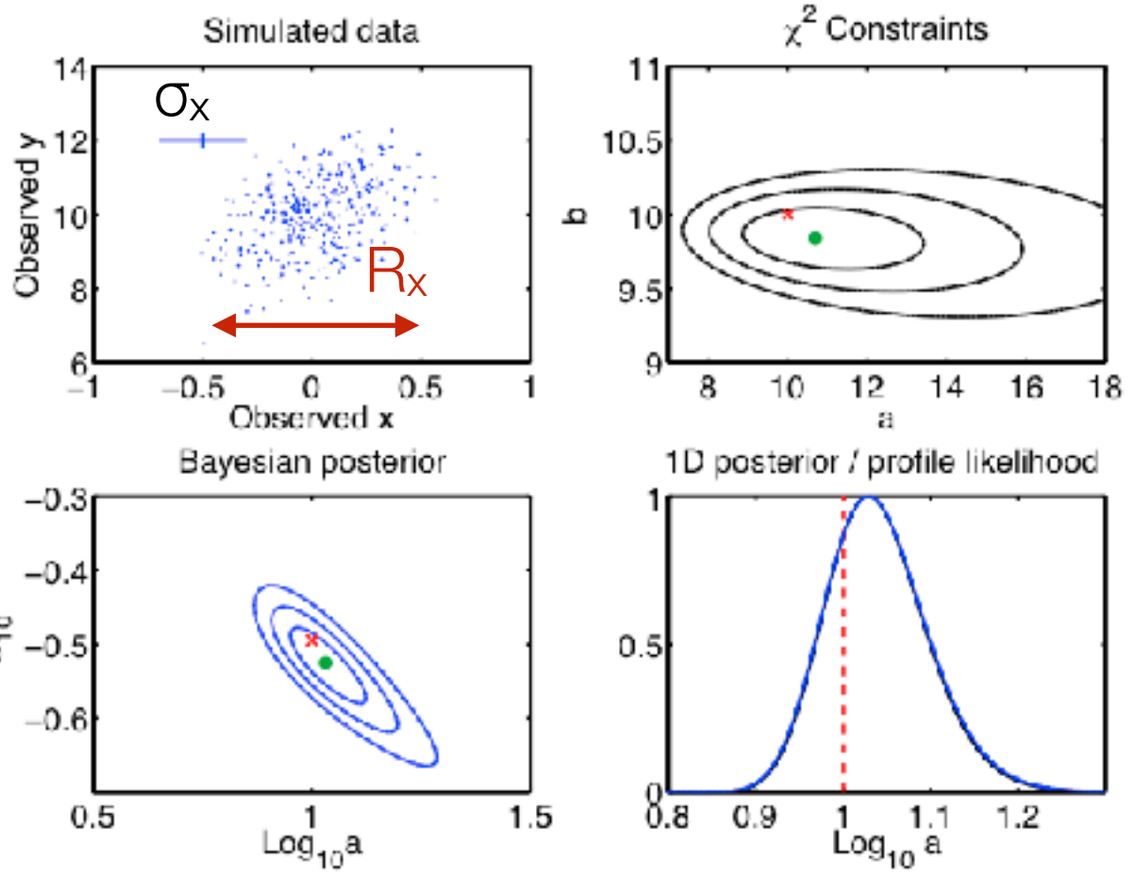
- An **observed** x value far from the origin is more probable to arise from **up-scattering of a lower latent x value** (due to noise) than down-scattering of a higher (less frequent) x value

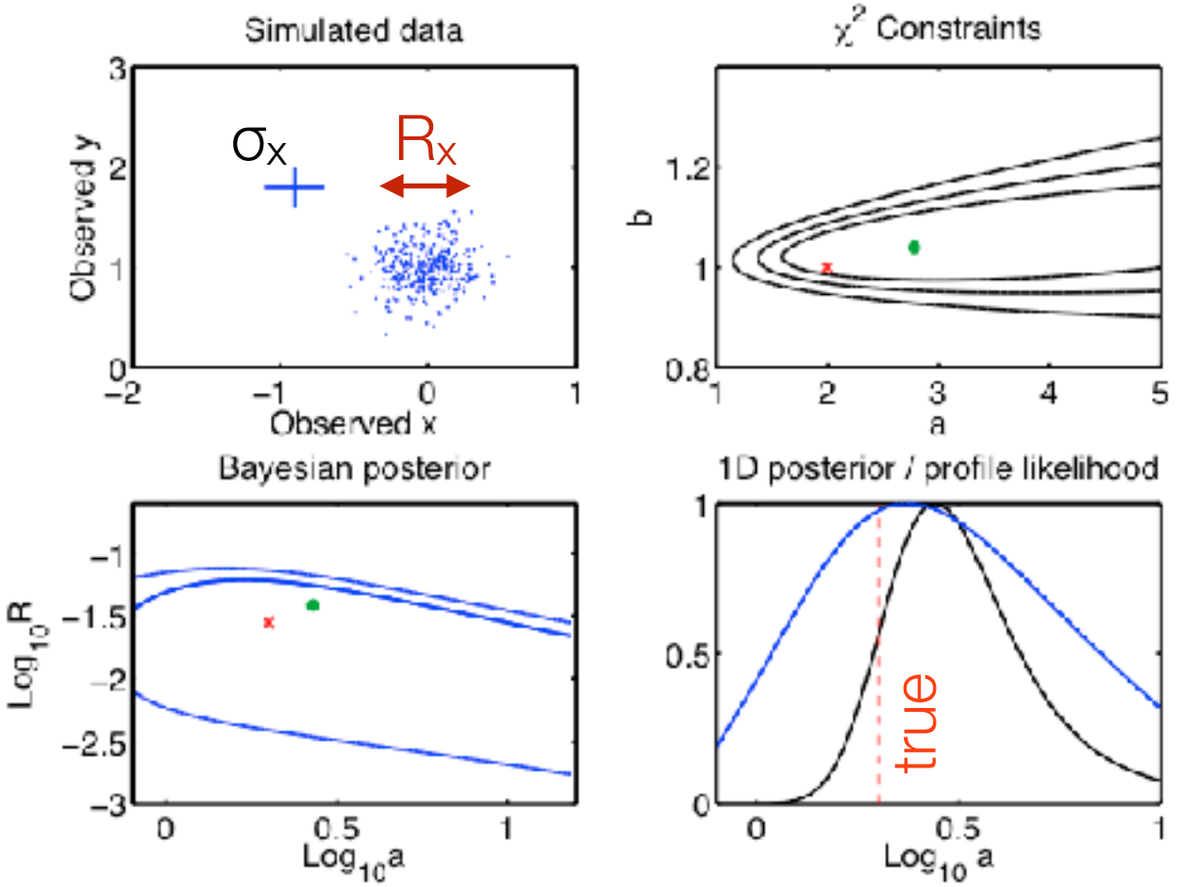
The key parameter is noise (σ_x) to population (R_x) characteristic variability scale ratio

$\sigma_x/R_x << 1$

$$y_i = b + ax_i$$

$\sigma_x/R_x \sim 1$





Bayesian (black) marginal posterior identical to Chi-Squared (blue)

Bayesian marginal posterior broader but less biased than Chi-Squared

March, RT et al (2011)

Slope reconstruction

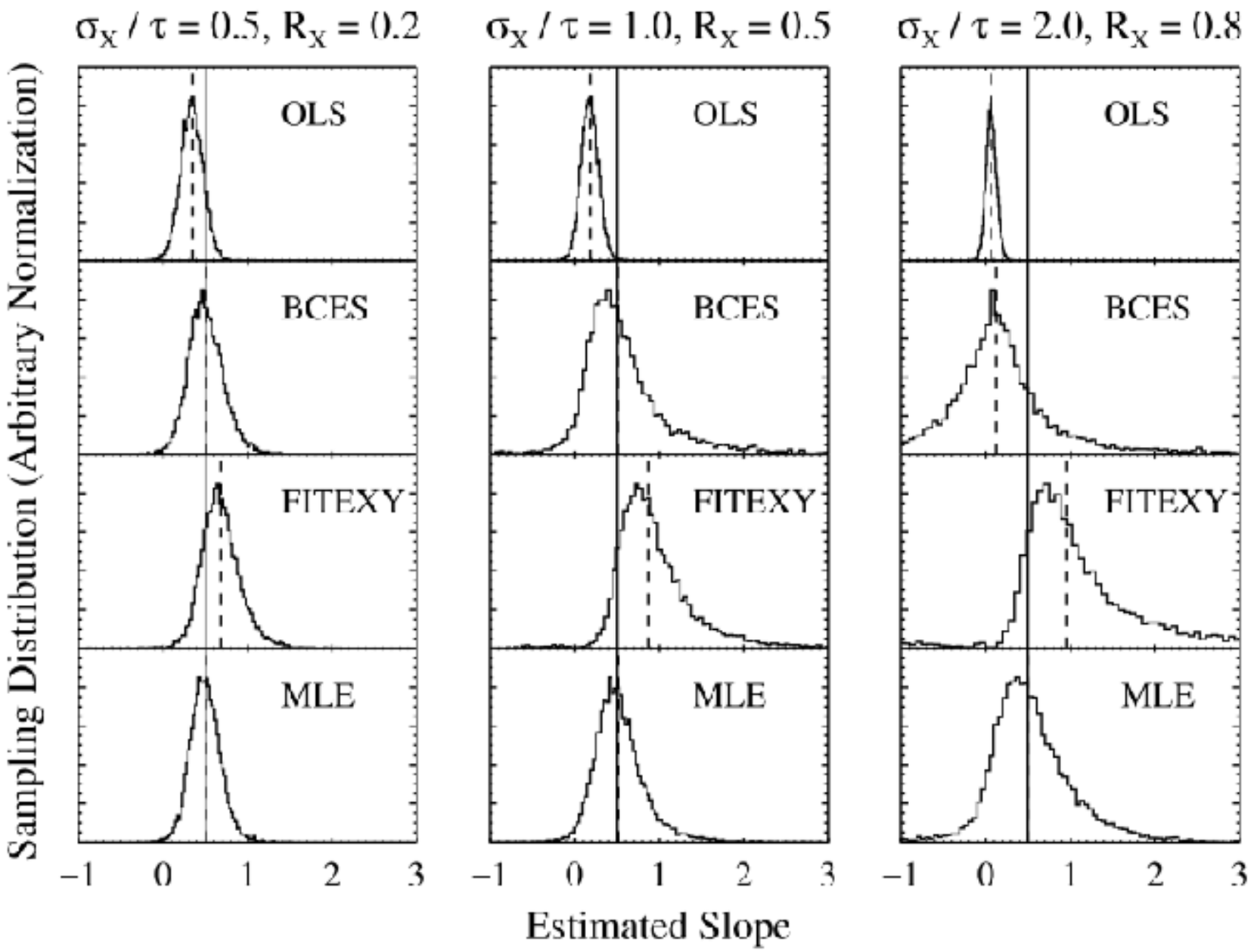
Imperial College London

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

Ordinary Least Square

Chi-Square incl variance

Maximum Likelihood



BIASSED LOW

BIASSED HIGH

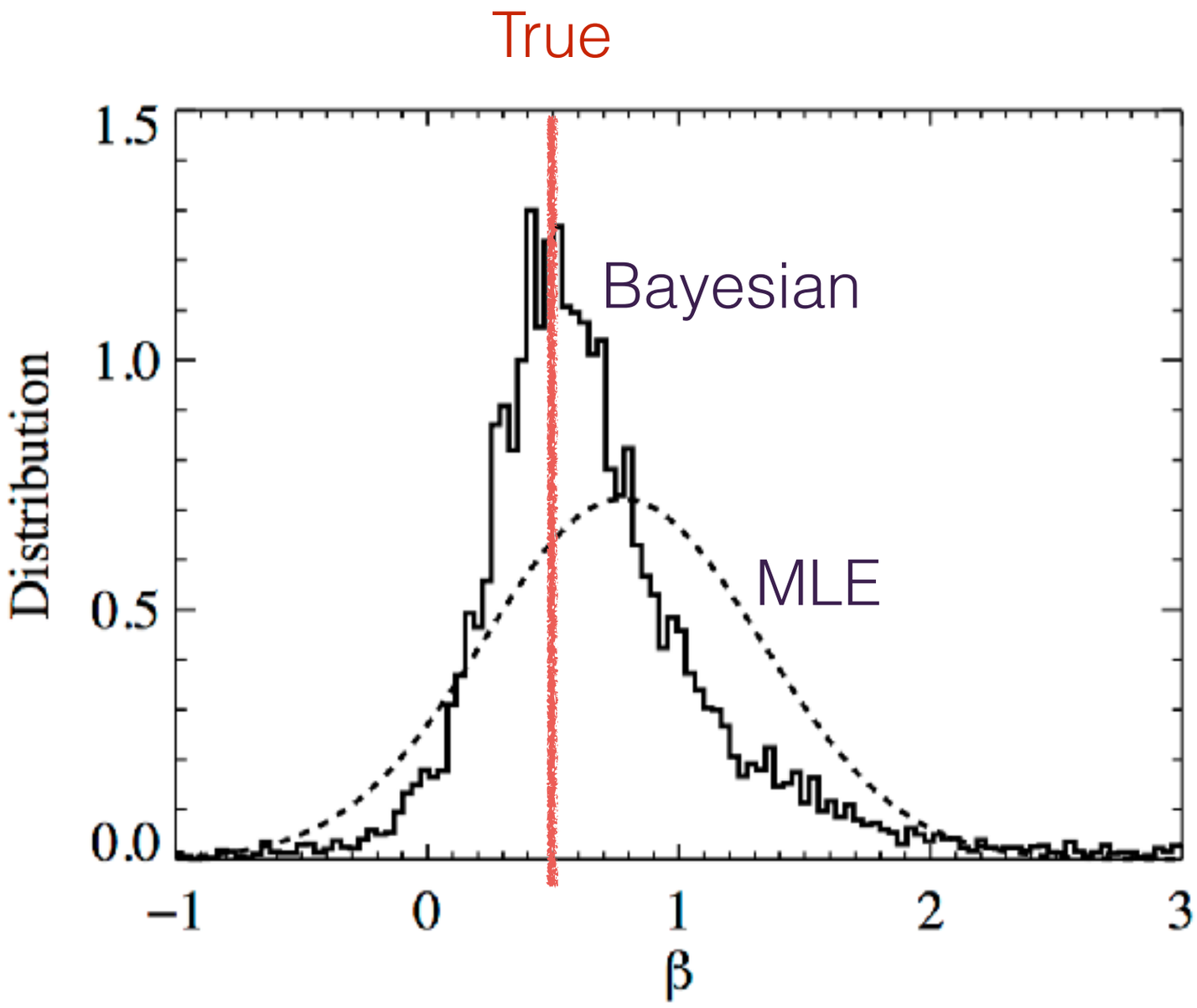
APPROX UNBIASSED

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

Why should you care?

Imperial College London

$R_x = \sigma_x^2/\mathrm{Var}(x) = 1$ in this example: Comparing the MLE (dashed) with the Bayesian Hierarchical Model Posterior (histogram)



Standard MLE (or Least Squares/ Chi-Squared) fits are biased!

(even if you artificially inflate the errors to get Chi-Squared/dof ~ 1)

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

Supernovae Type Ia Cosmology example

Imperial College London

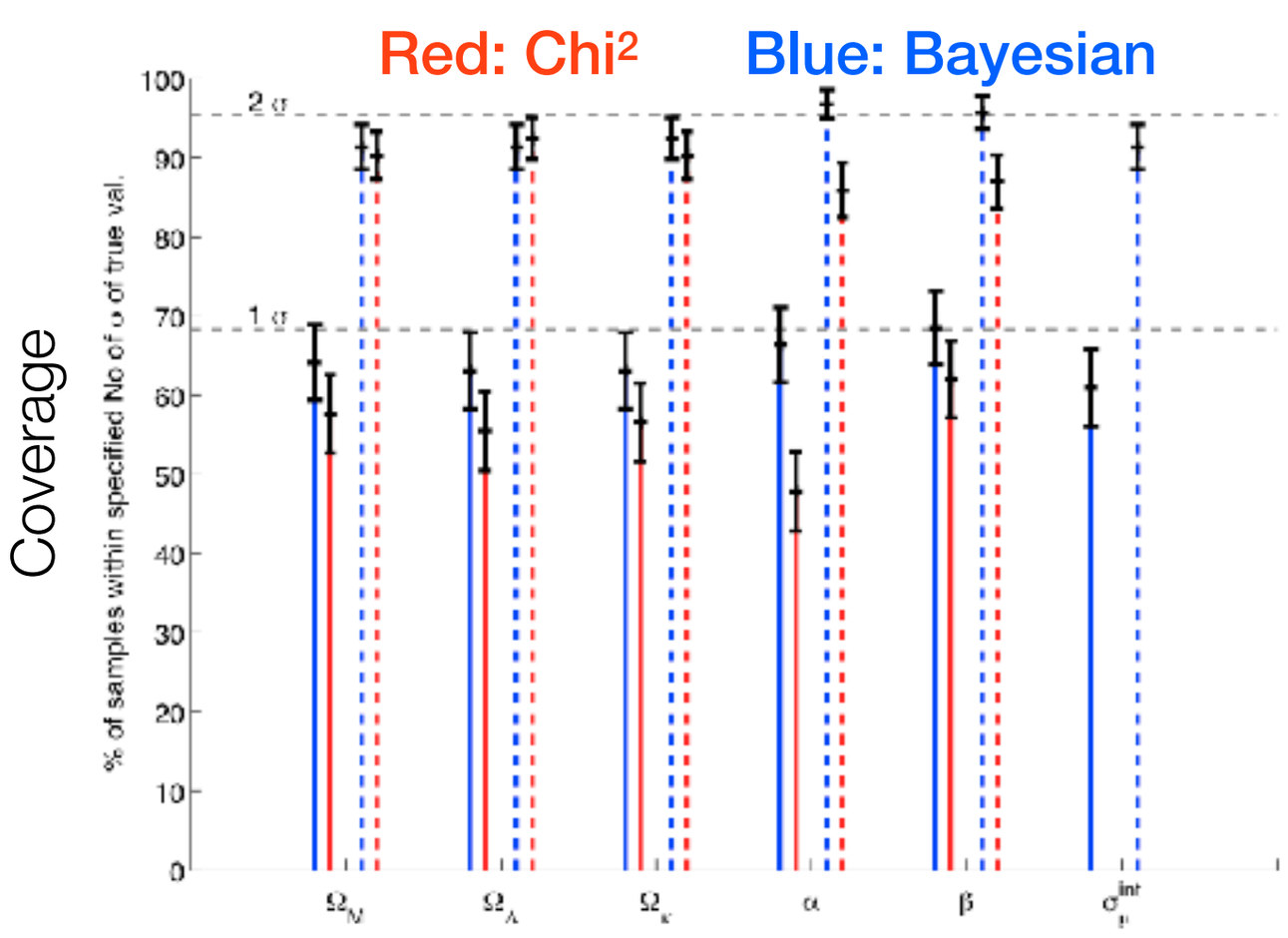
- Coverage of Bayesian 1D marginal posterior CR and of 1D Chi² profile likelihood CI computed from 100 realizations

- 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 (Chi²).



Results:

Coverage: generally improved (but still some undercoverage observed)

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

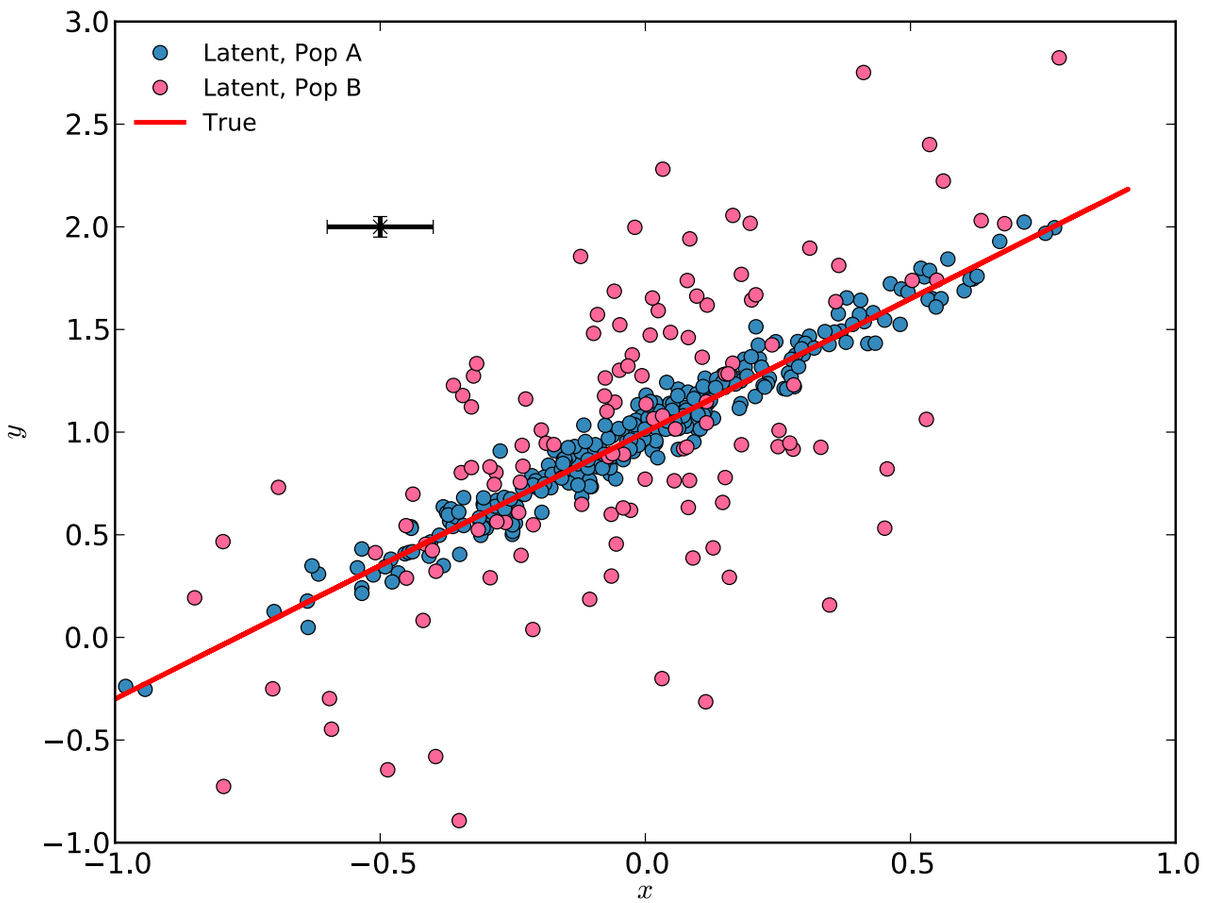
MSE: reduced by a factor 1.5-3.0 for all parameters

Adding object-by-object classification

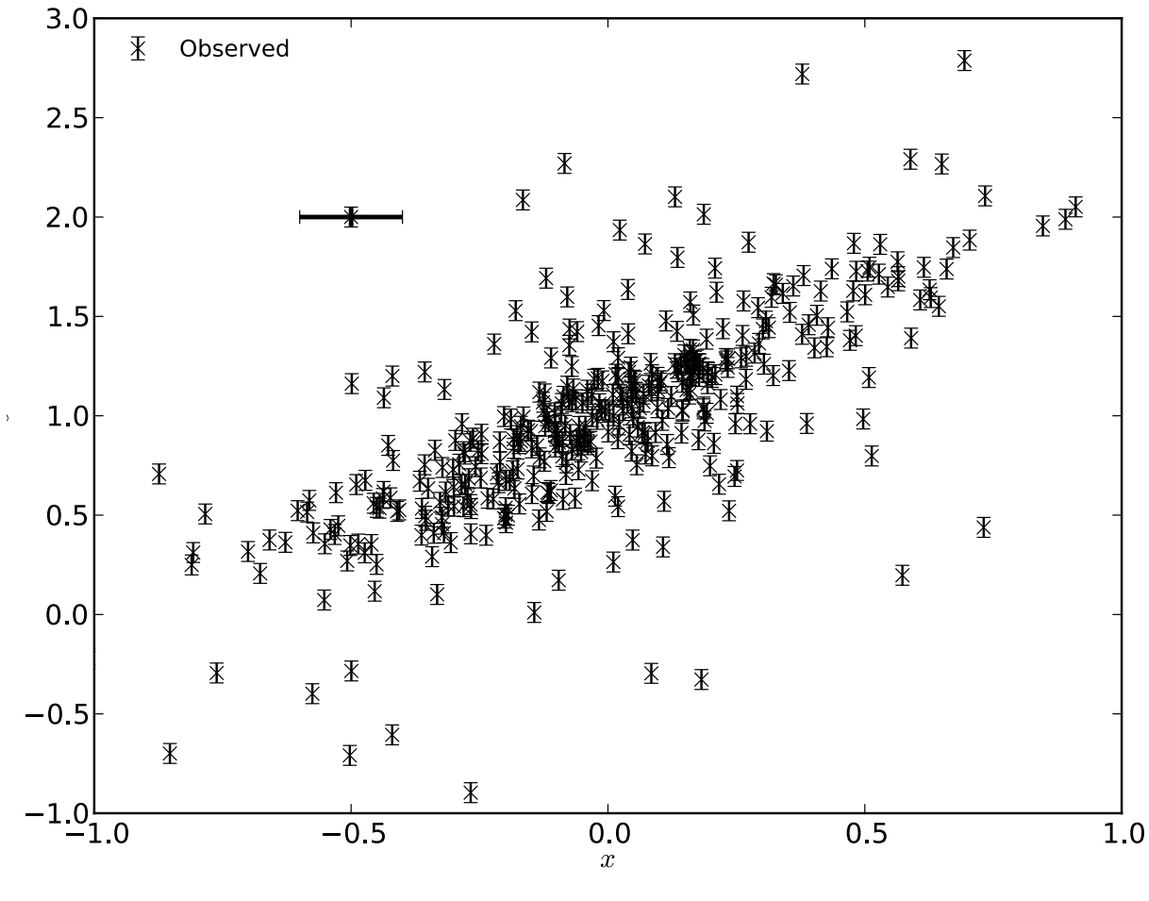
Imperial College London

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

LATENT



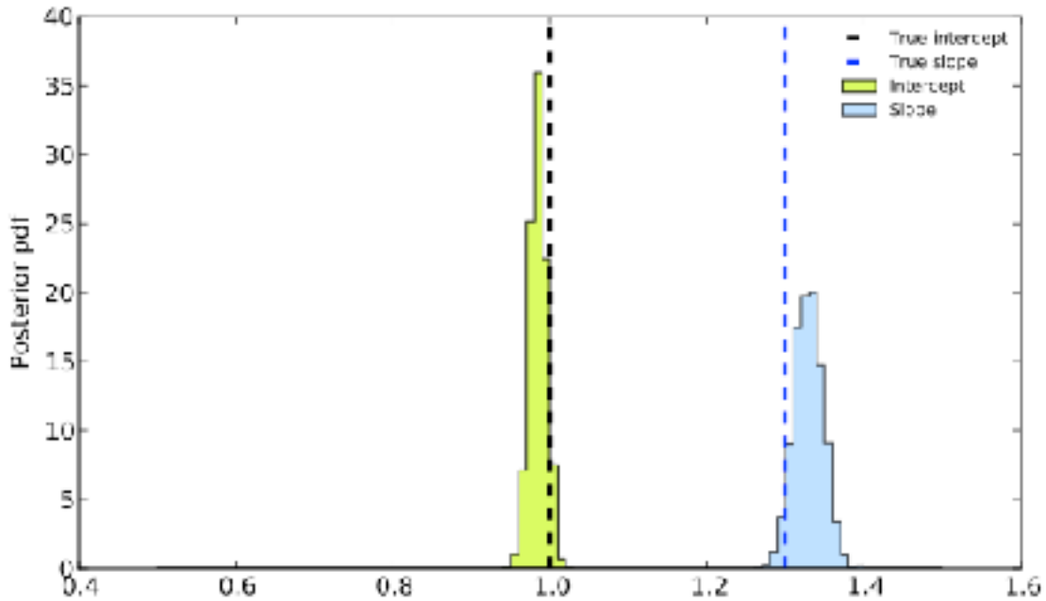
OBSERVED

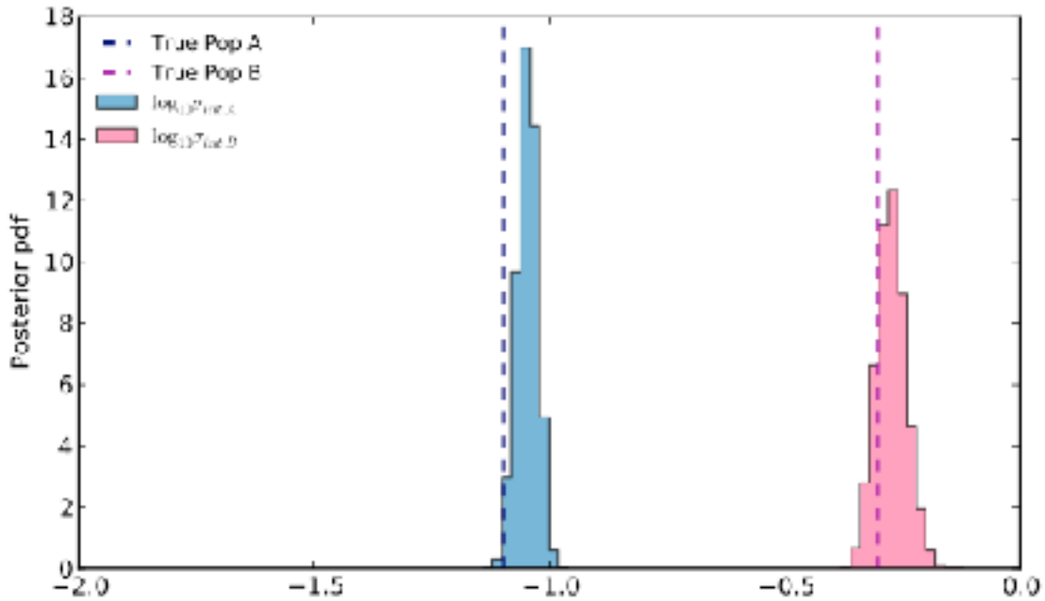


Reconstruction (N=400)

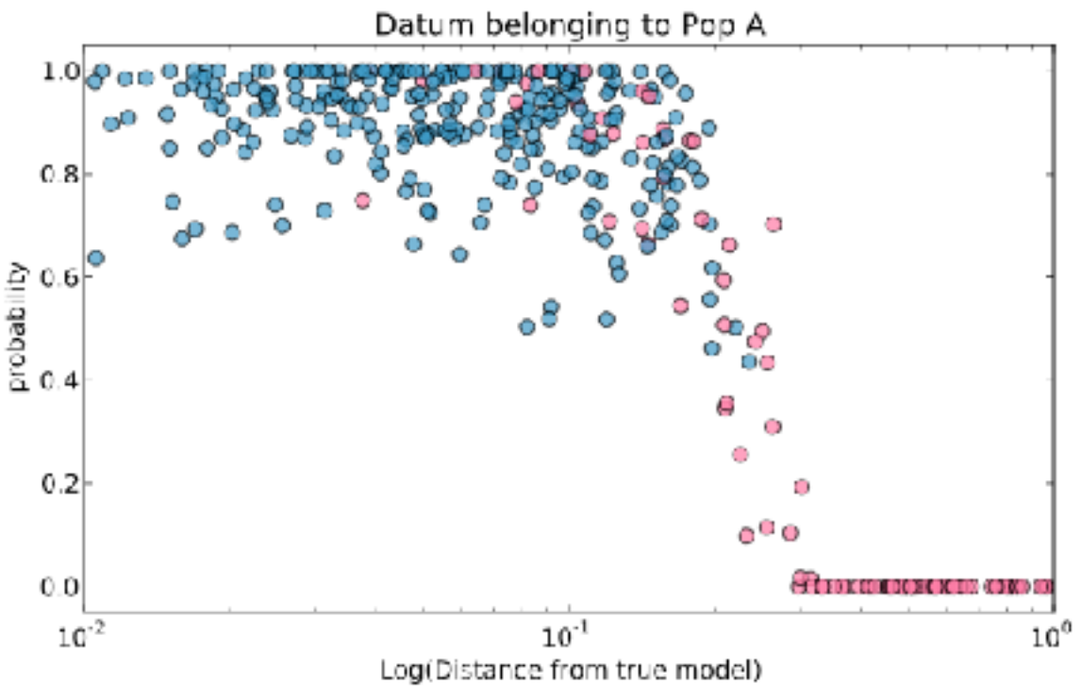
Imperial College London

Parameters of interest





Classification of objects



Population-level properties