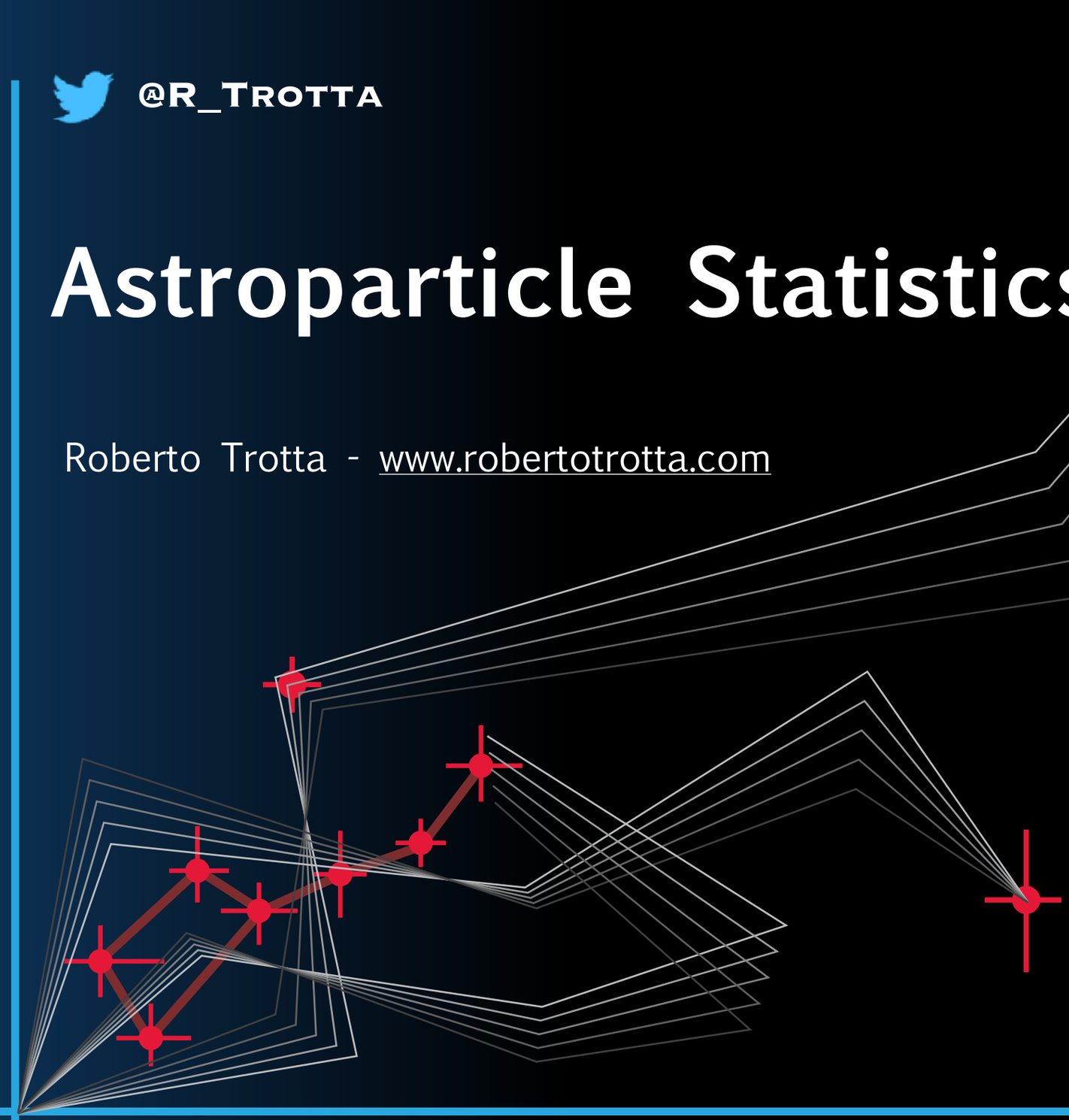
@R_Trotta

Astroparticle Statistics

Roberto Trotta - www.robertotrotta.com



The Dark and Visible Side of the Universe
Summer School, Texel, NL, July 4/5 2017

ICIC
Imperial Centre
for Inference & Cosmology

Imperial College
London

Get in teams of 4

Imperial College London

- Rules:
  - No more than 2 people from the same Institution
  - At least 1 woman / at least 1 man
  - At most 1 non-student
  - Choose a name for your team (a constellation)

Counts data

Imperial College London

- Example of data:
  - Number of counts for a source over area A for integration time t
  - As above, divided in energy bins ("binned data”)
  - As above, with an energy value for each count (“unbinned data”)
  - Additionally: energy and location measurement errors
  - Additionally: in the presence of a background
  - Additionally (and usually): background uncertainty

Poisson processes

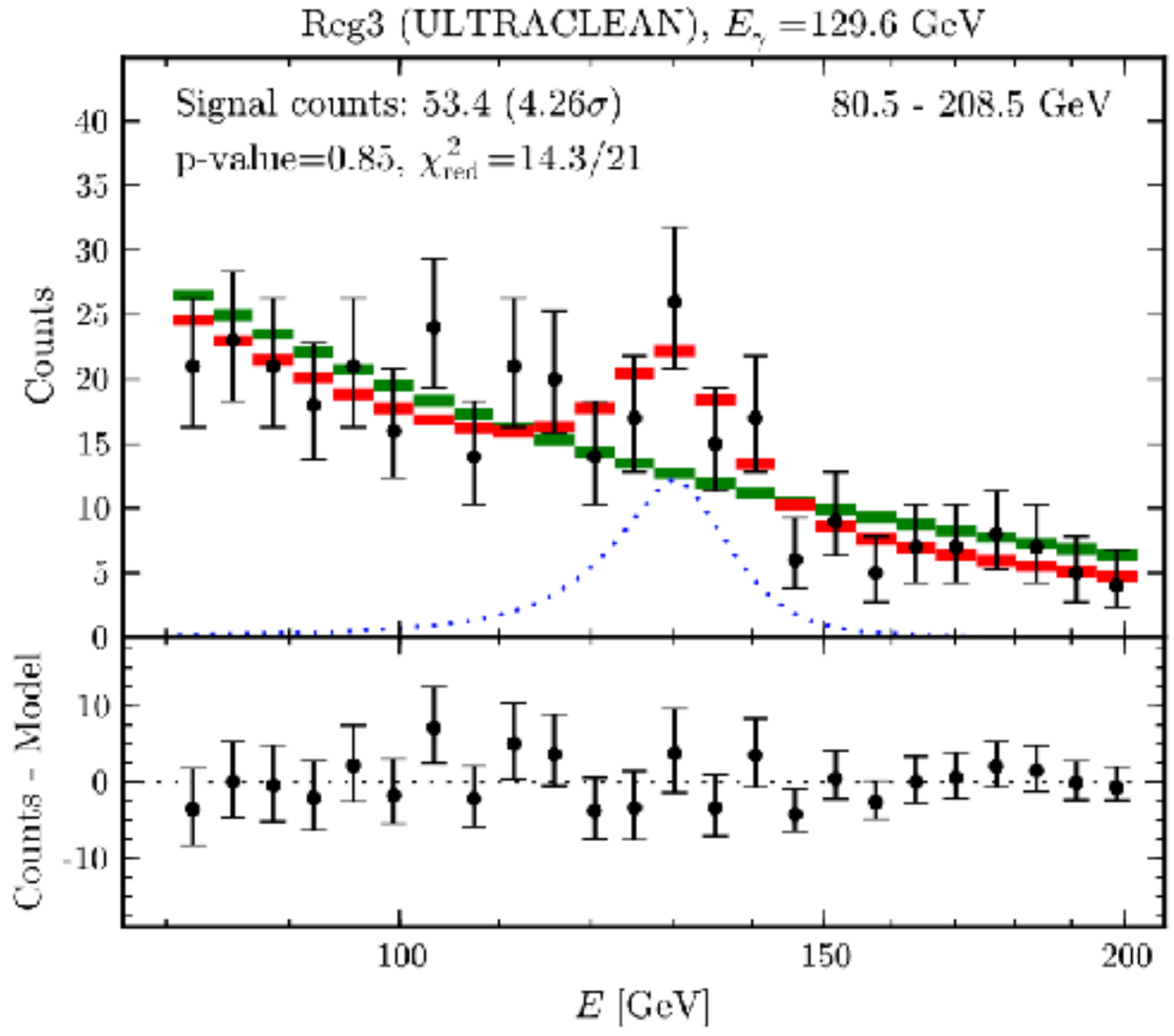
Imperial College London

- **Poisson model:** Used to model count data (i.e., data with a discrete, non-negative number of counts).
- Dark-matter related examples:
  - Gamma-ray photons
  - Dark matter direct detection experiments
  - Neutrino observatories
  - Collider data

Gamma-ray example

Imperial College London

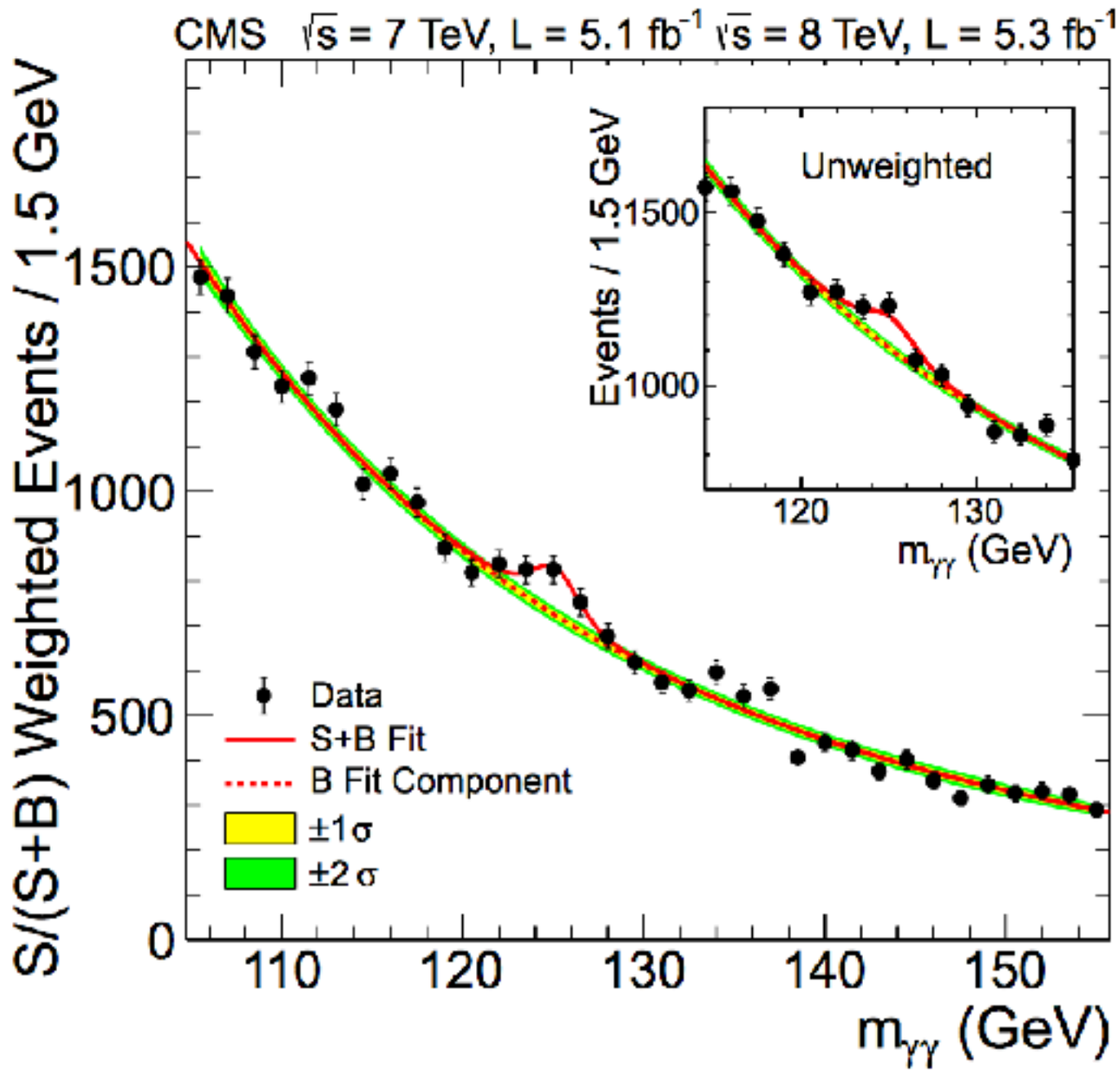
C. Weniger,
“A Tentative Gamma-Ray Line from Dark Matter Annihilation at the Fermi Large Area Telescope”, JCAP 1208 (2012) 007



LHC example

Imperial College London

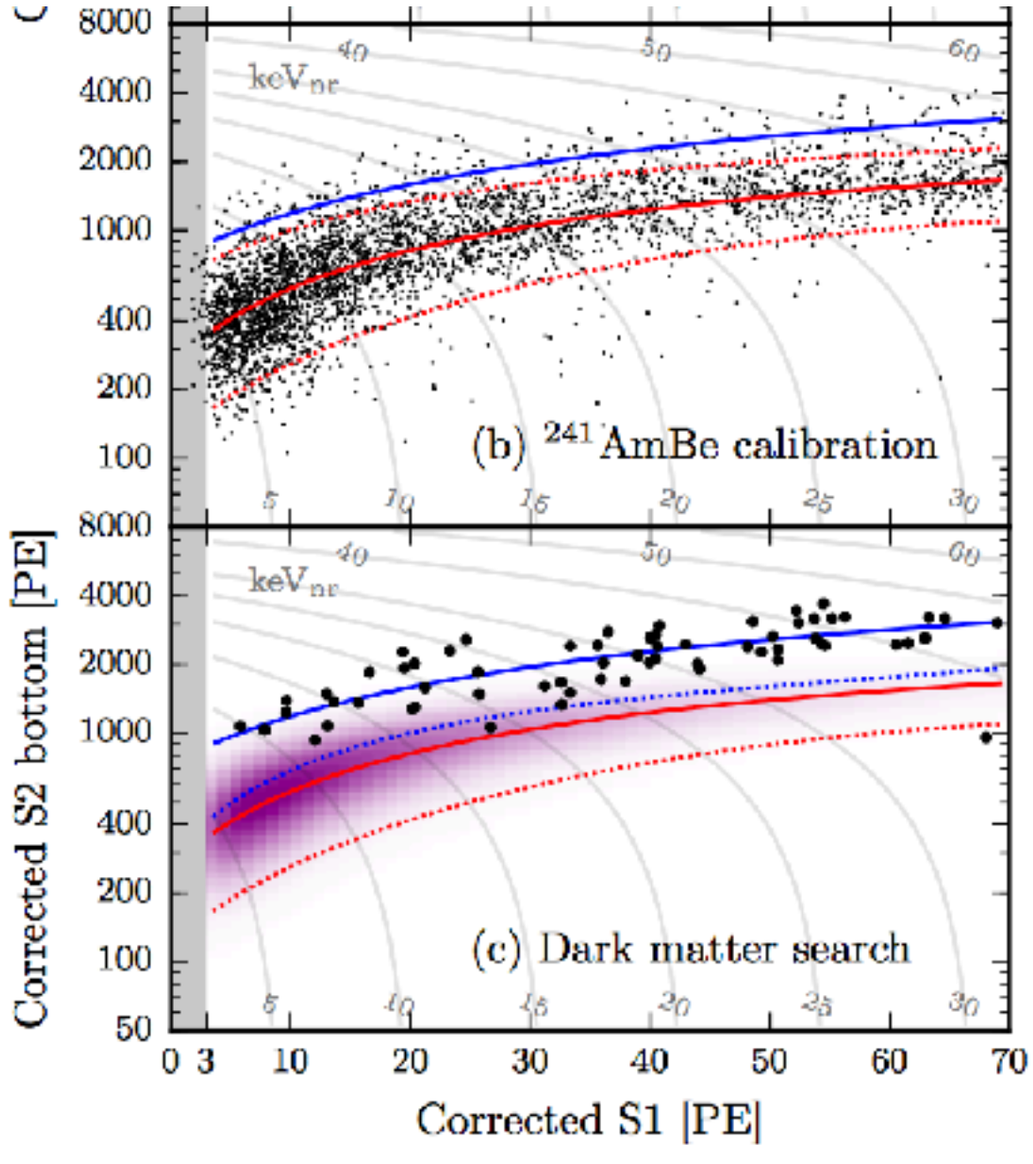
CMS Collaboration, Physics Letters B, 716, 30-61 (2012)

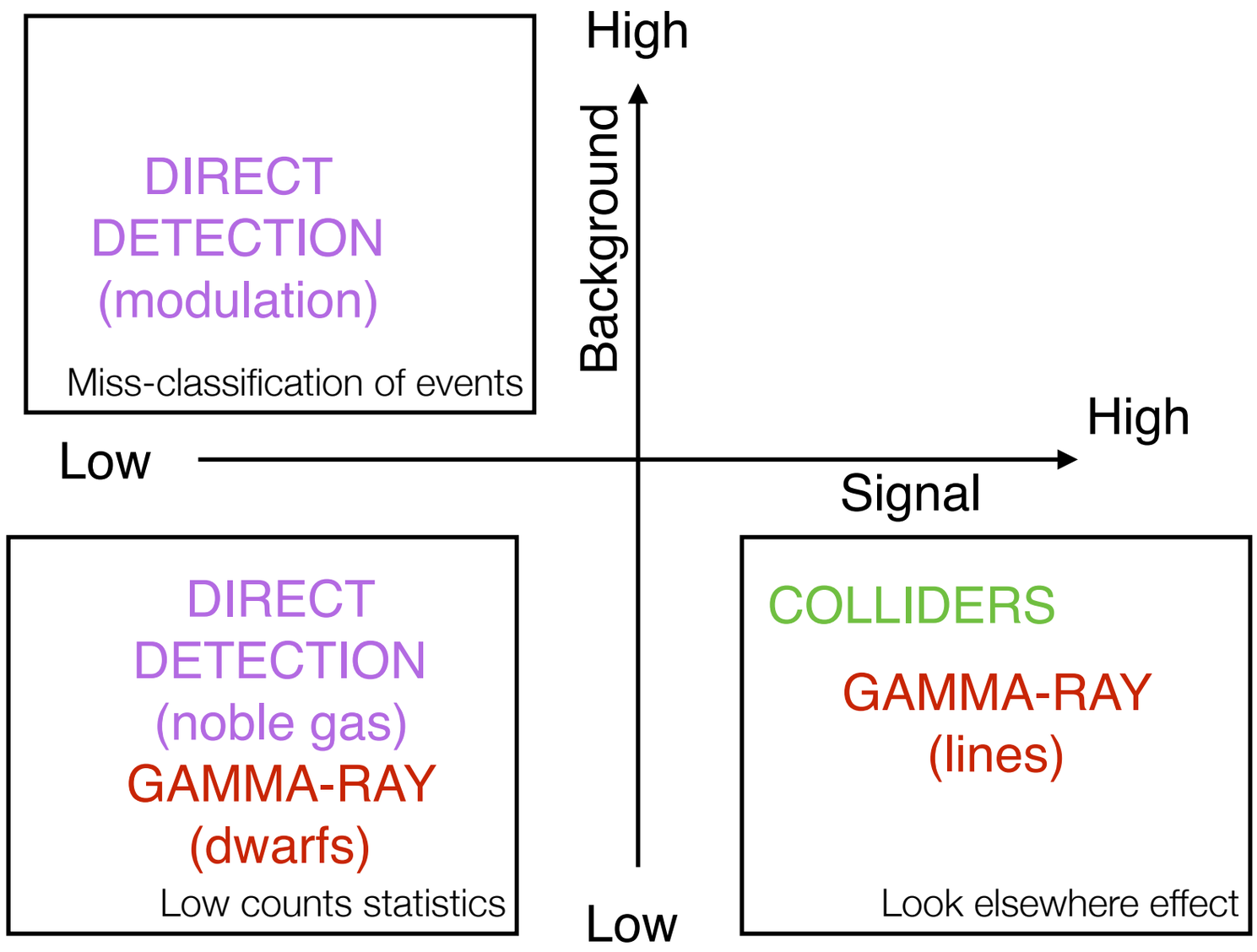


Direct detection example

Imperial College London

Aprile et al (Xenon1 Collaboration), “First Dark Matter Search Results from the XENON1T Experiment”, arxiv: 1705.06655




DIRECT
DETECTION
(modulation)
Miss-classification of events
High
Background
Low
High
Signal
DIRECT
DETECTION
(noble gas)
GAMMA-RAY
(dwarfs)
Low counts statistics
Low
COLLIDERS
GAMMA-RAY
(lines)
Look elsewhere effect

Poisson distribution

Imperial College London

- The **Poisson distribution** describes the probability of obtaining a certain number of events in a process where events occur with a fixed average rate and independently of each other:

$$P(r|\lambda, t) \equiv \text{Poisson}(\lambda) = \frac{(\lambda t)^r}{r!} e^{-\lambda t}.$$

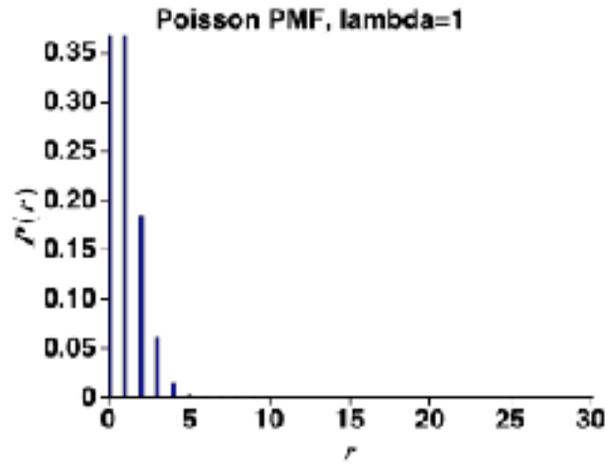
Data: r (counts)

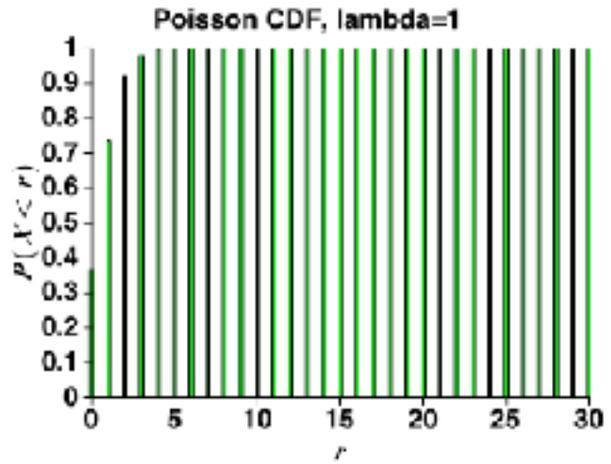
Parameter: λ (average rate, units 1/time. Often, t=1, i.e., choose the right units for λ)

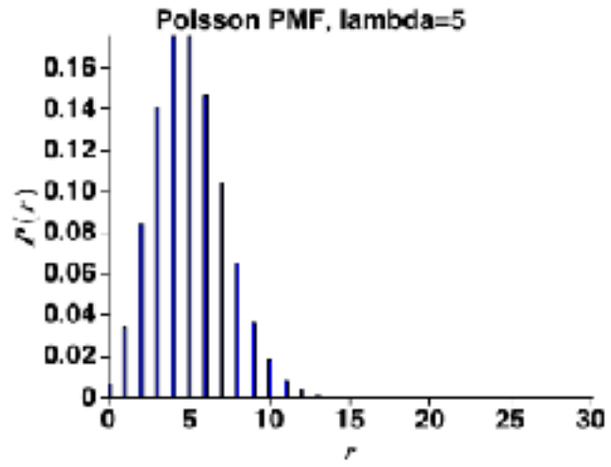
$$E(X) = \lambda t \qquad\qquad \text{Var}(X) = \lambda t$$

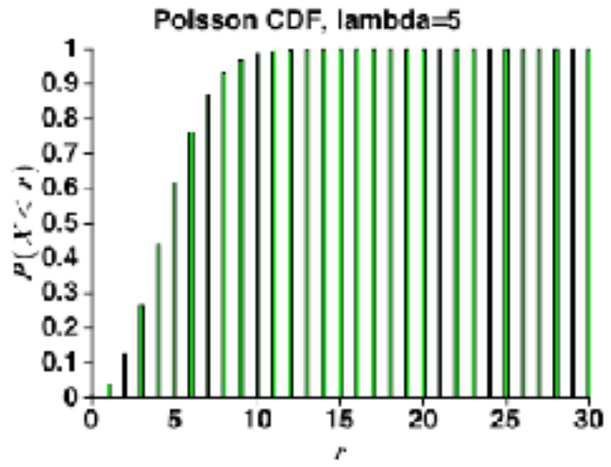
Probability Mass Function

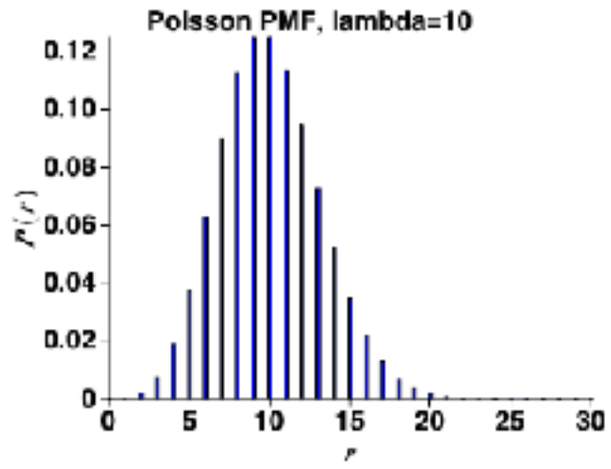
Cumulative Distribution Function

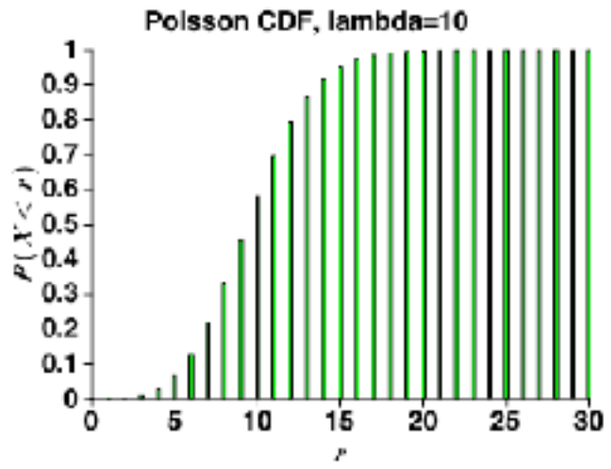




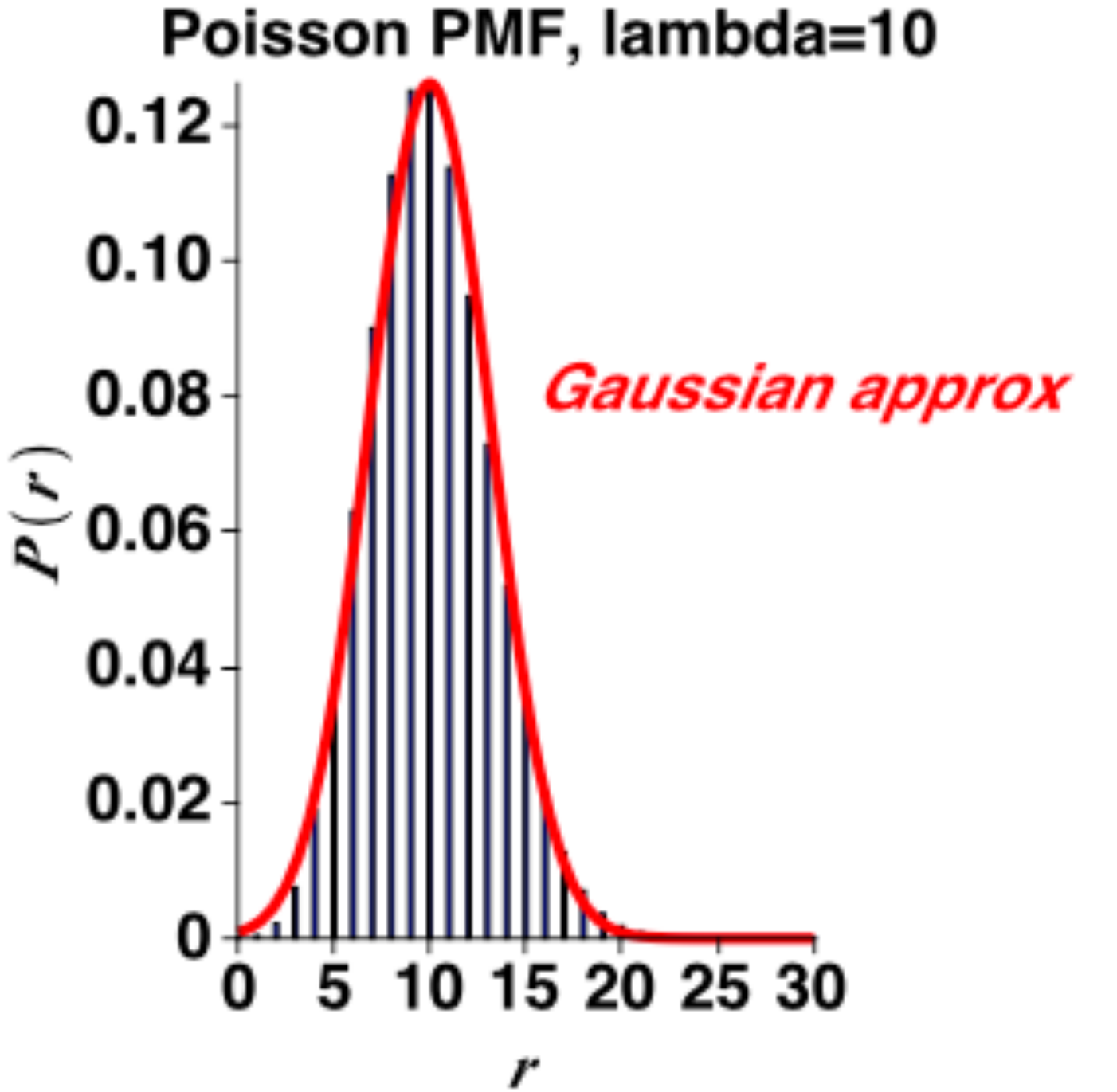








Gaussian approximation for lambda >> 1.
Beware: Gaussian is a continuous pdf!



Inference and the likelihood function
(on the board)

1. Distributions and random variables
2. The likelihood function
3. The inferential problem
4. The Maximum Likelihood Estimator (MLE) and its meaning

Confidence Intervals (tricky!)

Imperial College London

- **Frequentist confidence intervals:** give the probability of observing the data (over an ensemble of measurements) given the true value of the parameter.

- This is usually **not** what you are interested in! (i.e., the probability for the value of the parameter which is a Bayesian statement!)

- Let θ be the parameter of interest; $[\theta_1, \theta_2]$ the confidence interval, which is a function of the measured value, x (both x and $[\theta_1, \theta_2]$ vary across repeated experiments).

- The confidence interval $[\theta_1, \theta_2]$ is a member of a set (over repeated experiments, **with fixed θ**) such that **the set** has the property that:

$$P(\theta \in [\theta_1, \theta_2]) = \alpha$$

  for every allowable θ. If this is true, the set of intervals is said to have “correct coverage”.

- **The set of intervals** contains the true θ with probability α. **IMPORTANT: It does not mean that $[\theta_1, \theta_2]$ contains the true θ with probability α !**

Confidence Belt Construction

Imperial College London

- The classic Neyman construction for Confidence Intervals is as follows:

1. For each possible value of θ, draw horizontal acceptance interval with the property that:
$$P(x \in [x_1, x_2]|\mu) = \alpha$$
2. Measure x and obtain the value $\hat{x}$
3. Draw a vertical line at the observed value of x
4. The α confidence interval [θ_1, θ_2] is the union of all values of θ for which the acceptance interval is intercepted by the vertical line

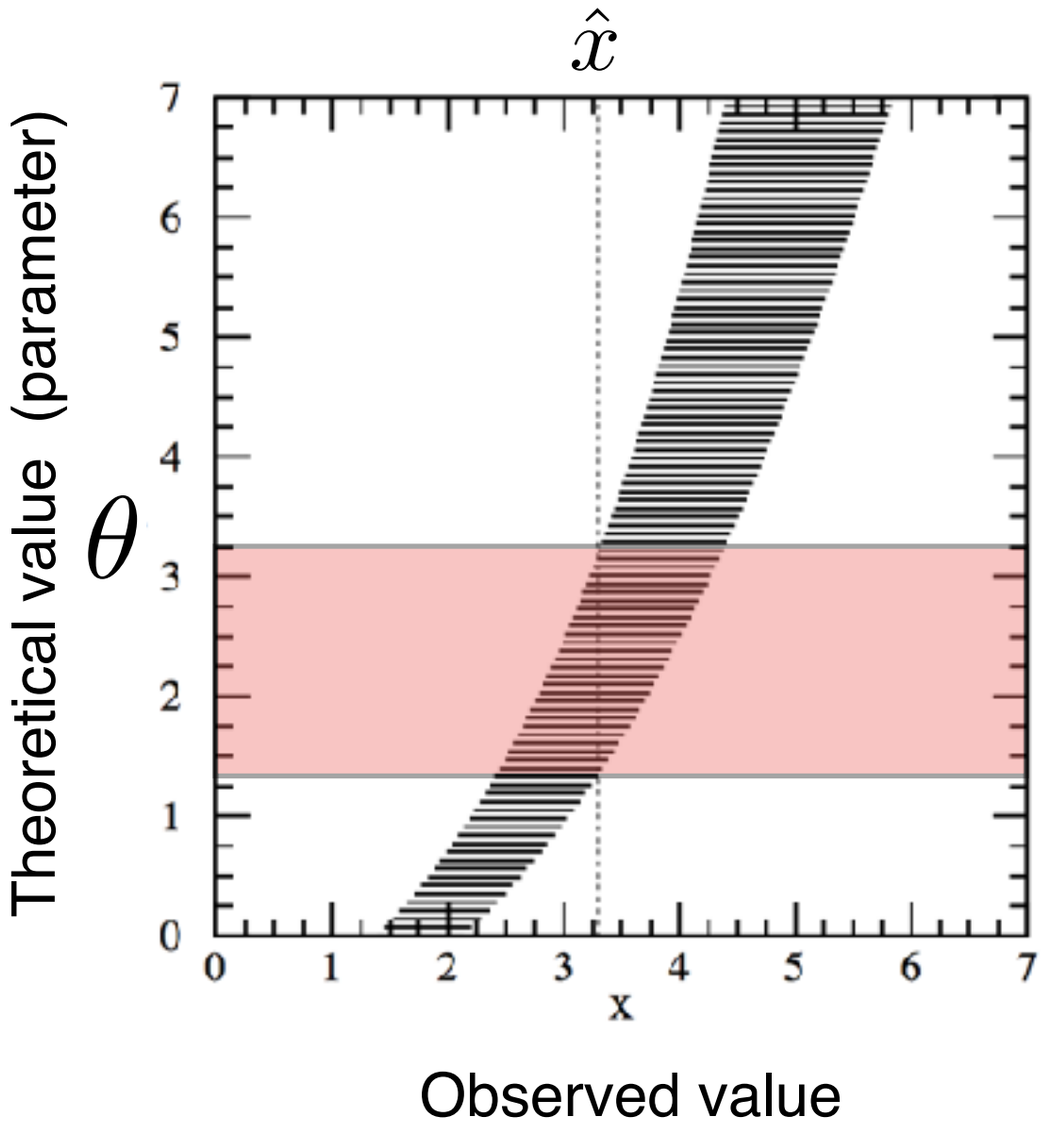


Figure from Feldman & Cousins (1997)

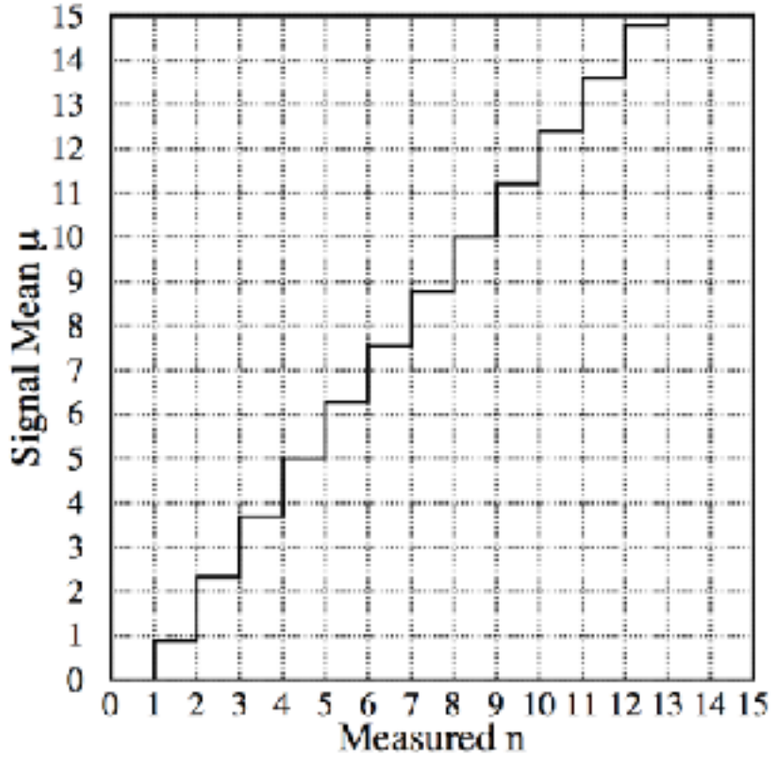
Confidence Belt for Poisson counts

Imperial College London

- **Problem:** To define the acceptance region uniquely requires the specification of auxiliary criteria (arbitrary choice).
- Choice 1 (“upper limits”): $P(x < x_1|\theta) = 1 - \alpha \Rightarrow P(\theta > \theta_2) = 1 - \alpha$
- Choice 2 (“central confidence intervals”):

$$P(x < x_1|\theta)P(x > x_2|\theta) = (1 - \alpha)/2 \Rightarrow P(\theta < \theta_1)P(\theta > \theta_2) = (1 - \alpha)/2$$

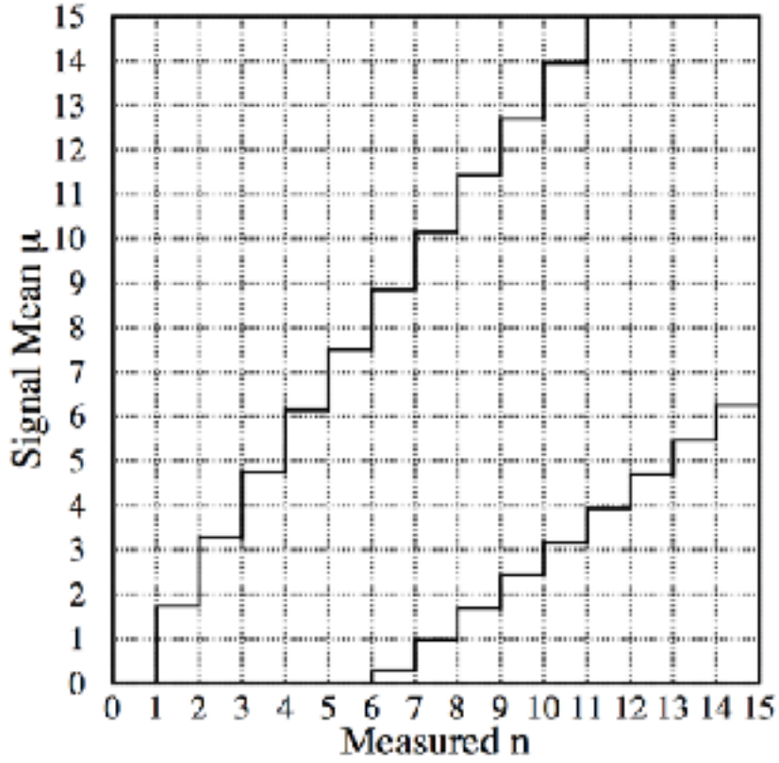
90% upper C.L.



b = 3 (known)

n = s+b

90% central C.L.



Figures from Feldman & Cousins (1997)

The Flip-Flopping Physicist

Imperial College London

- “I’m doing an experiment to look for a New Physics signal over a known background. But obviously I don’t know whether the signal is there. So I can’t choose whether to go for an upper limit or a 2-sided confidence interval ahead of time”

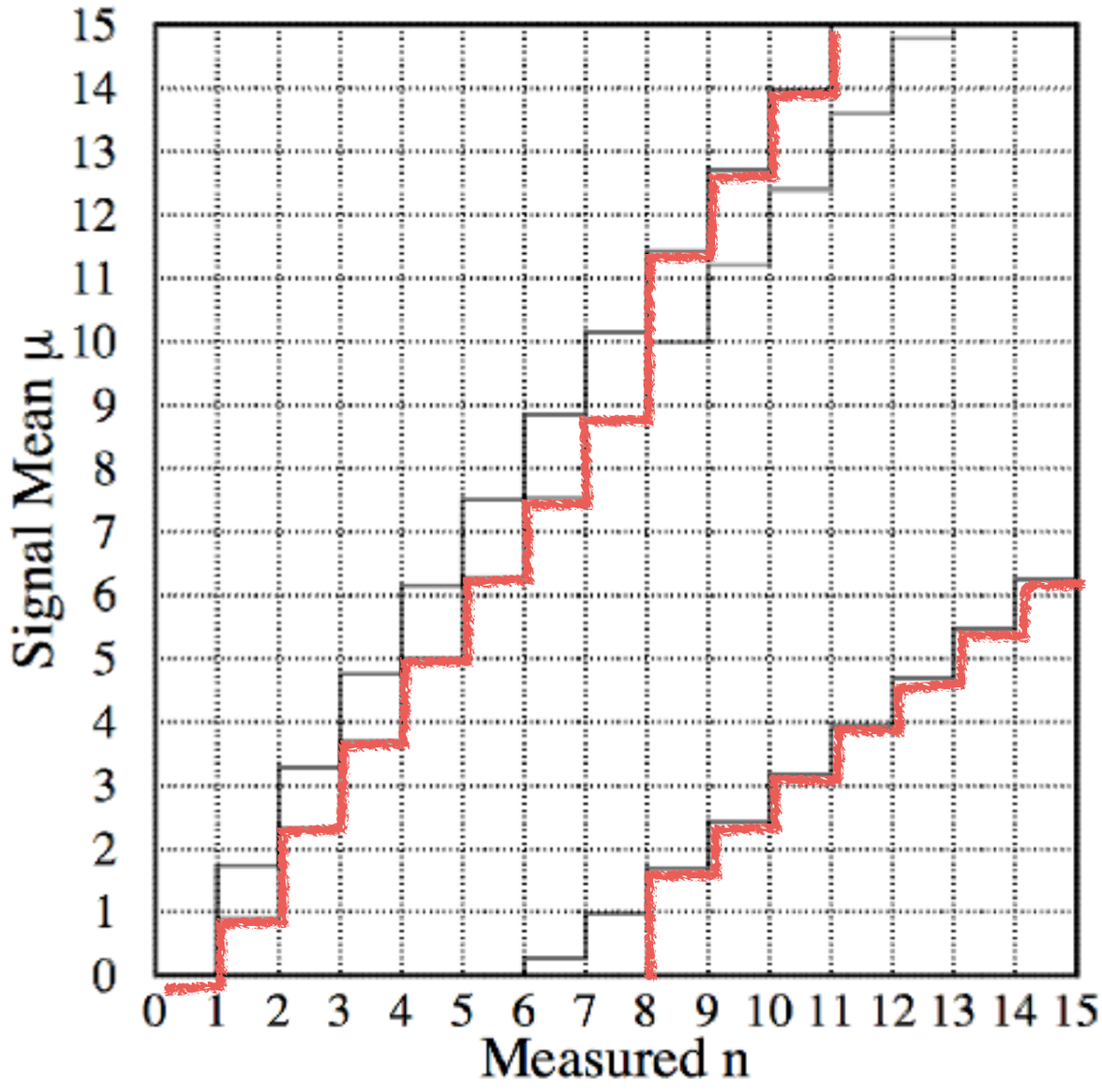
- “Here’s what I’m going to do: if my measurement x is less than 3σ away from 0, I’ll quote upper limits only; if it’s more than 3σ away from 0, I’ll quote a 2-sided interval (discovery). And if x is < 0, I’ll pretend that I’ve measured 0 to build in some conservatism”

- **Problem:** Your choice for the Confidence Belt construction now depends on the data! This leads to breaking its coverage properties.

Flip-Flopping

Imperial College London

- **Problem:** The coverage property of the confidence belt is now lost



Adapted from Feldman & Cousins (1997)

Empty Confidence Intervals

Imperial College London

- **A further problem:** in the presence of a known background (b):
  - use confidence belt construction to determine confidence interval for s+b
  - subtract (known) b to get confidence interval for s
  - But if x<b (i.e., measured counts less than expected background), the confidence interval for signal is the empty set.

Feldman & Cousins Construction

Imperial College London

- The flip-flopping problem and the empty sets problem are solved by the Feldman & Cousins construction (arxiv: physics/9711021).

- In essence: An ordering principle, decides which values of x to include in the confidence belt based on the likelihood ratio (stop when C.L. α is reached):

$$R = P(n|\mu)/P(n|\mu_{\rm best})$$

- Here, μ_{best} is the (physical) value of the signal mean rate that maximises the probability of getting n counts, i.e., μ_{best} = max(0, n-b).

μ = 0.5, b = 3.0 here

Feldman & Cousins (1997)

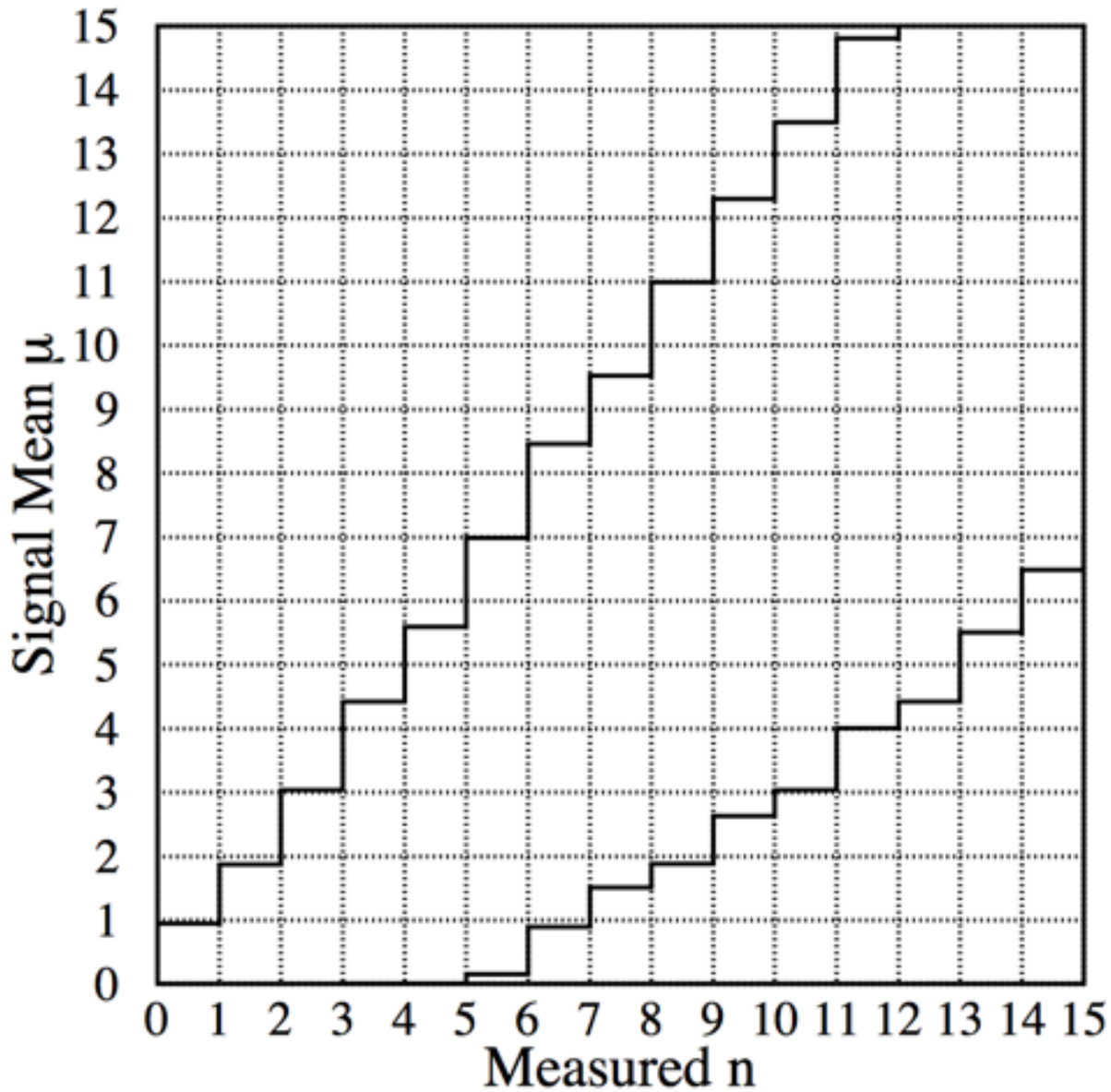
n	$P(n\|\mu)$	$\mu_{\rm best}$	$P(n\|\mu_{\rm best})$	R	rank	U.L.	central
0	0.030	0.	0.050	0.607	6		
1	0.106	0.	0.149	0.708	5	√	√
2	0.185	0.	0.224	0.826	3	√	√
3	0.216	0.	0.224	0.963	2	√	√
4	0.189	1.	0.195	0.966	1	√	√
5	0.132	2.	0.175	0.753	4	√	√
6	0.077	3.	0.161	0.480	7	√	√
7	0.039	4.	0.149	0.259		√	√
8	0.017	5.	0.140	0.121		√	
9	0.007	6.	0.132	0.050		√	

Feldman & Cousins Belt

Imperial College London

b = 3 (known)

n = s+b



Feldman & Cousins: Summary

Imperial College London

PROS

1. Guarantees coverage
2. Does an automatic flip-flopping (from 1-sided to 2-sided intervals) while preserving the probability content of the belt (horizontally!)
3. No empty sets
4. No unphysical values for the parameters

CONS

1. Can be complicated to construct
2. Difficult to extend to many-dimensions
3. Still suffers from a weird pathology (see below)

Experiment 1: b =0. n=0. s< 2.44 (90% CL)
Experiment 2: b =10. n=0. s< 0.93 (90% CL)

Exp 2 just got lucky! (downward fluctuations of larger b). Why should they report more stringent limits than Exp 1?!

High Level Summary #1

Imperial College London

- The likelihood function is NOT a pdf for the parameters. You cannot interpret it as such (this requires Bayes Theorem!).

- Even the simplest case of inferring the underlying mean rate of Poisson-distributed counts is fraught with difficulties in the classic Frequentist approach.

- MLE estimates and classic Neyman confidence intervals for the rate of a Poisson signal in the presence of a (known) background can (and do) give non-sensical results. What should you do in this case?

- Feldman & Cousins ordering principle fixes some of the pathologies (flip-flopping; negative estimates; empty intervals) but not all of them (downward fluctuations of large expected backgrounds lead to smaller upper limits: nonsensical).

- And: We haven't even talked about nuisance parameters (e.g., unknown b) yet!

- A possible solution: use Bayes theorem and forget about all of the above! (next up)

Bayesian Inference

Imperial College London

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

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \text{(Bayes Theorem)}$$

A simple rule with profound consequences!

A → **θ:** parameters
B → **d:** data

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

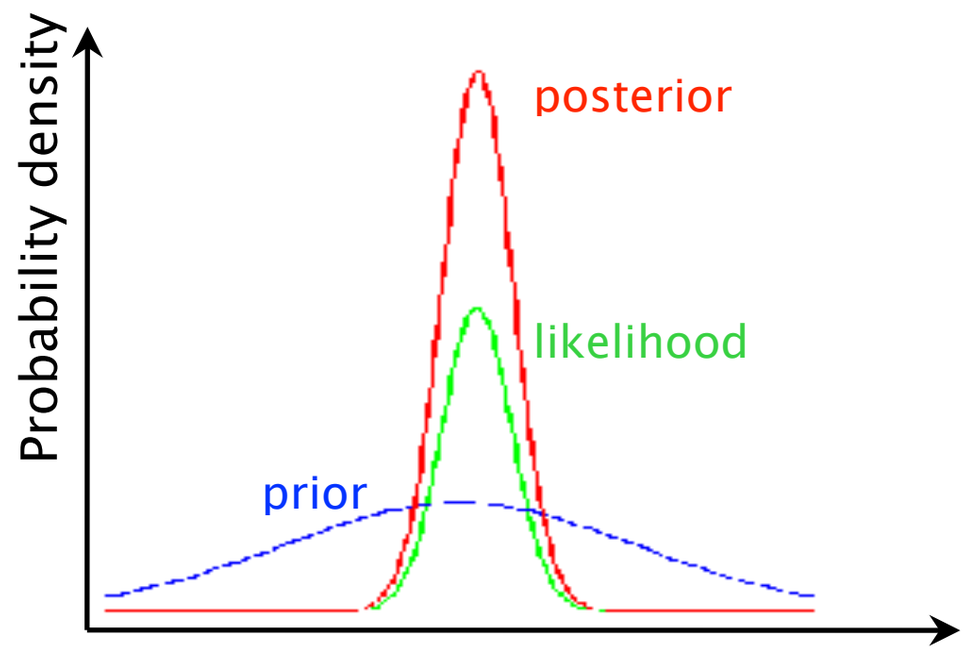
$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Bayesian Inference

Imperial College London

posterior likelihood prior

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



The posterior is a pdf, the likelihood isn't!
Posterior ≠ likelihood, even for uniform priors
"Non-informative" priors don't exist
Beware "uniform" priors (i.e. p(θ) = const)

Posterior ≠ likelihood

Imperial College London

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

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

P(hypothesis|data) ≠ P(data|hypothesis)

Hypoth

Data: th

P(F|

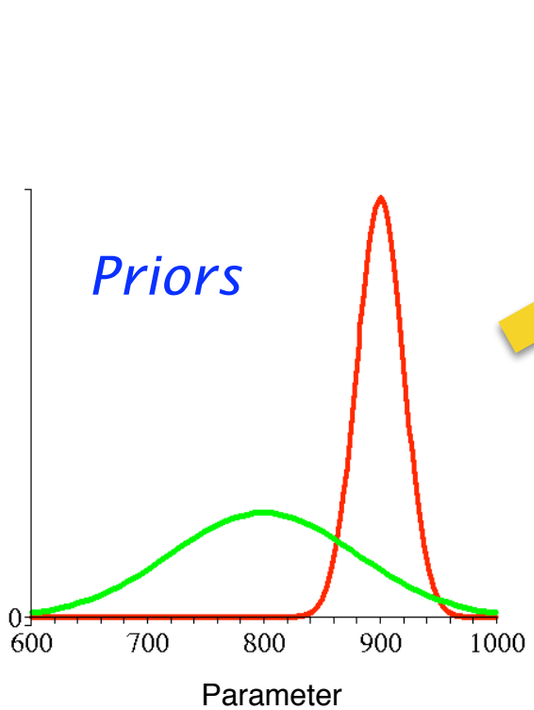
0.03

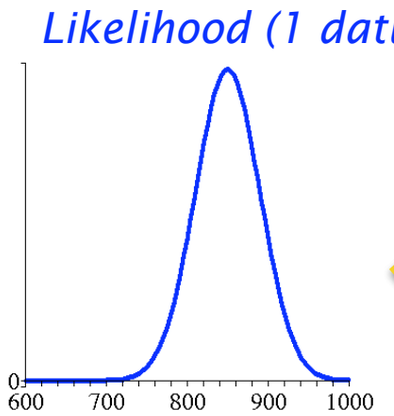
> *"Bayesians address the question everyone is interested in by using assumptions no–one believes [the prior], while Frequentists use impeccable logic to deal with an issue of no interest to anyone [the data distribution]"*
> (Louis Lyons)

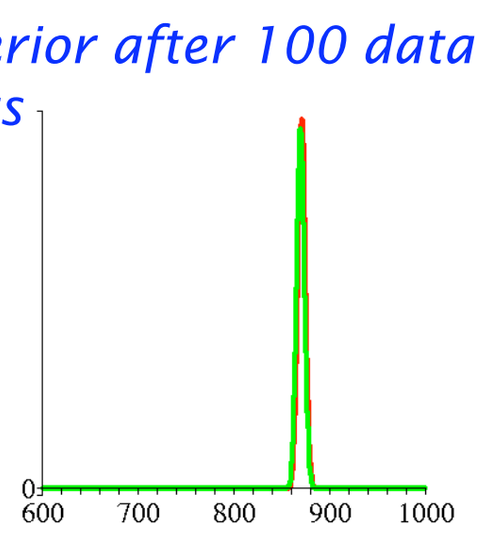
The Matter with Priors

Imperial College London

The prior dependence will in principle vanish for strongly constraining data (not true for model selection!)
In practice: we always work at the sensitivity limit





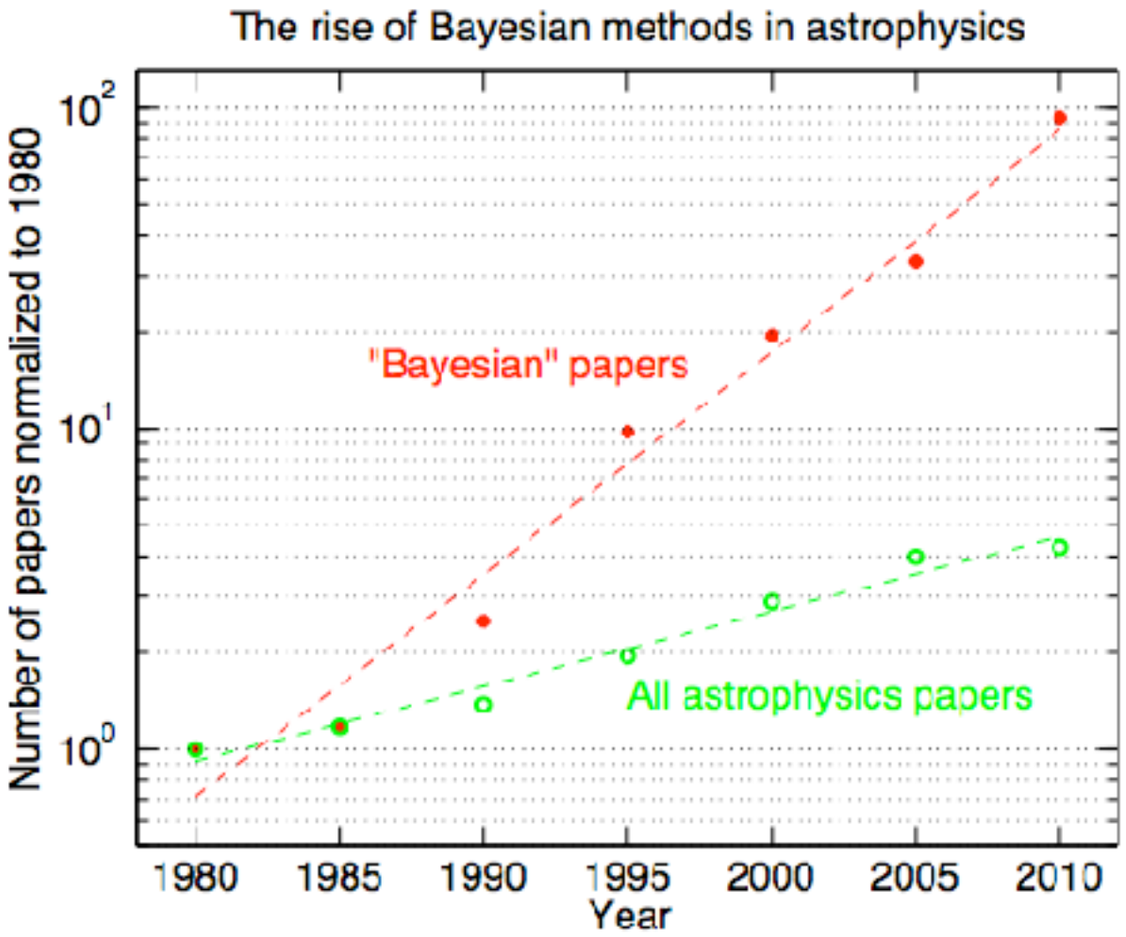


Scientist A

Scientist B

Bayesian Methods on the Rise

Imperial College London



Advantages

- Principled: it gives the the quantity you are after
- Scalable: e.g. MCMC w 10^5 parameters
- Extremely simple marginalization
- Accommodates model complexity
- Efficient (for expensive likelihoods)
- Easy to use: Metropolis-Hastings is ~6 lines of pseudocode
- Comprehensive inferential framework:
  - parameter inference
  - model selection
  - model averaging
  - classification
  - prediction
  - optimisation

Review article: RT (2008)

Reasons to be Bayesian

Imperial College London

- **Philosophy:** it gives the quantity you are after (the probability for the hypothesis) without invoking any ad hoc reasoning, defining test statistics, etc. Immune from stopping rule problems, and no ambiguity due to the set of imaginary repetition of data (frequentist). Inferences are conditional on the data you got (not on long-term frequency properties of data you will never have!)

- **Simplicity:** uninteresting (but important) nuisance parameters (e.g., instrumental calibration, unknown background) are integrated out from the posterior (marginalized) with no extra effort. Their uncertainty is fully and automatically propagated to the parameters of interest.

- **Insight:** The prior forces the user to think about their state of knowledge/ assumptions. Frequentist results often match Bayesian results with special choices of priors. (“there is no inference w/o assumptions”!).

- **Efficiency:** evaluating the posterior distribution in high-dimensional parameter spaces (via MCMC) is fast and efficient (up to millions of dimensions!).

Universal Bayesian inference guide

Imperial College London

- With Bayes, no matter what inferential problem you are facing, **you always follow the same (principled) steps:**

1. Identify **the parameters** in the problem, θ
2. Assign a **prior pdf** to them, based on your state of knowledge, $p(\theta)$. Make sure the prior is proper. Beware of uniform priors!
3. Write down **the likelihood function** (including as appropriate: backgrounds, instrumental effects, selection effects, etc), $P(d|\theta) = L(\theta)$
4. Write down the (unnormalized) **posterior pdf:** $P(\theta|d) \propto L(\theta)\, p(\theta)$
5. Evaluate the posterior, usually numerically (e.g. **MCMC sampling,** see later)
6. Report inferences on the parameter(s) of interest by showing **marginal posterior pdf's** (see later). The posterior pdf encodes the full degree of belief in the parameters values post data. It **is** a probability distribution **for** the parameters! (unlike any frequentist quantity!)

Bayesian Credible Intervals

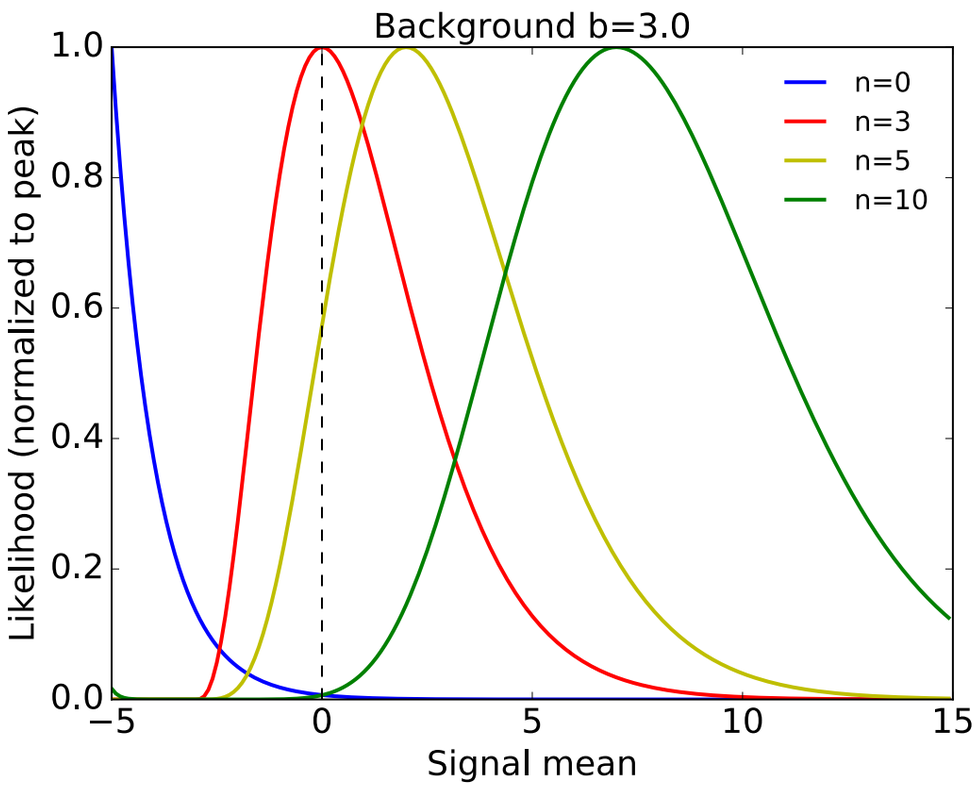
Imperial College London

- For a Bayesian, the posterior pdf describes the **full result of the inference.**
- It gives your state of knowledge after you have seen the data (updating your degree of belief encoded in the prior)
- You can quote α% Credible Intervals (NOT Confidence Intervals!) by giving any range of the posterior that contains α% of the probability. No need to give the mean +/- standard deviation in a Gaussian approximation!
- 3 obvious choices (make sure you say what you are picking!):
  - **Symmetric credible intervals:** (1-α)/2 % probability in each tail outside the interval
  - **Upper/Lower limits:** (1-α)% in the tail outside the interval
  - **Highest Posterior Density (HPD) intervals:** Come down from the peak of the posterior until you encompass α% of the probability. By construction also the shortest intervals.

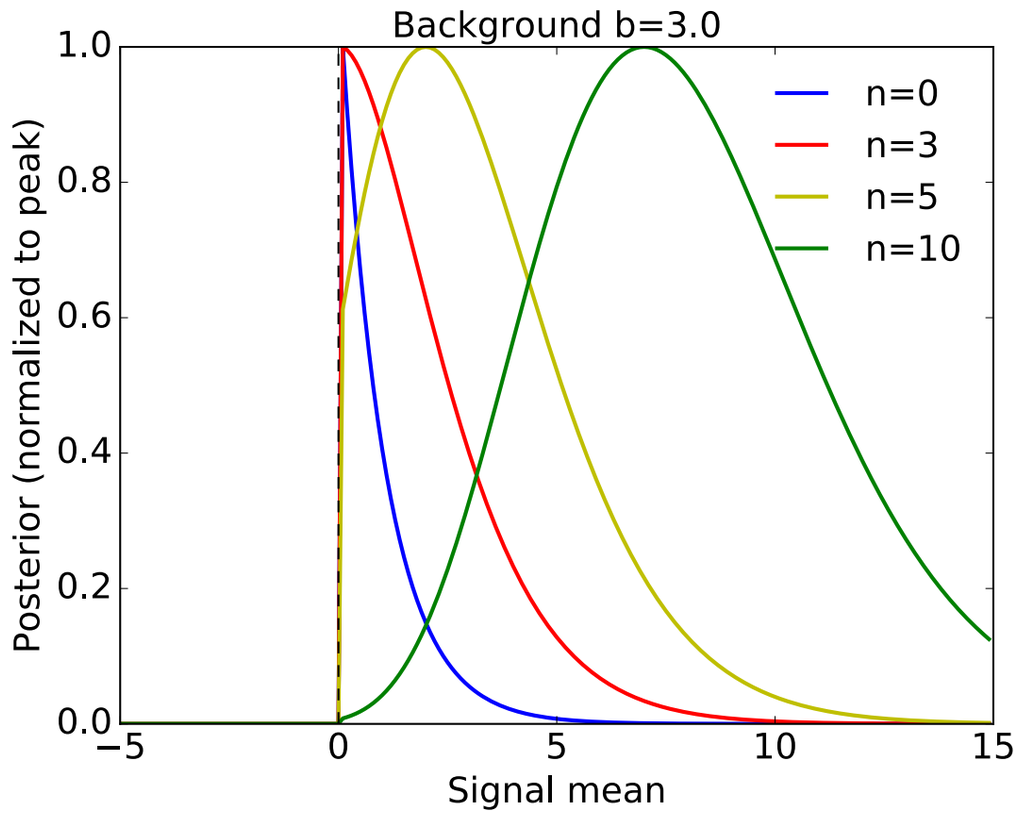
Example: counts experiment with Bayes (no background)

The power of priors

Likelihood



Posterior



(Improper) Prior = 1, if signal > 0
= 0, otherwise

The prior effortlessly incorporates information about the physical conditions for s
It removes the unphysical (s<0) region, which requires complicated shenanigans for Frequentists!

Inference in many dimensions

Imperial College London

Usually our parameter space is multi-dimensional: how should we report inferences for one (or 2) parameter at the time?

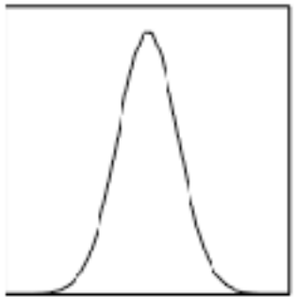
BAYESIAN: Marginal posterior (integration)	FREQUENTIST: Profile likelihood (maximisation)
$p(\theta\|d) = \int d\psi p(\theta, \psi\|d)$	$\mathcal{L}(\theta) = \max_\psi \mathcal{L}(\theta, \psi)$

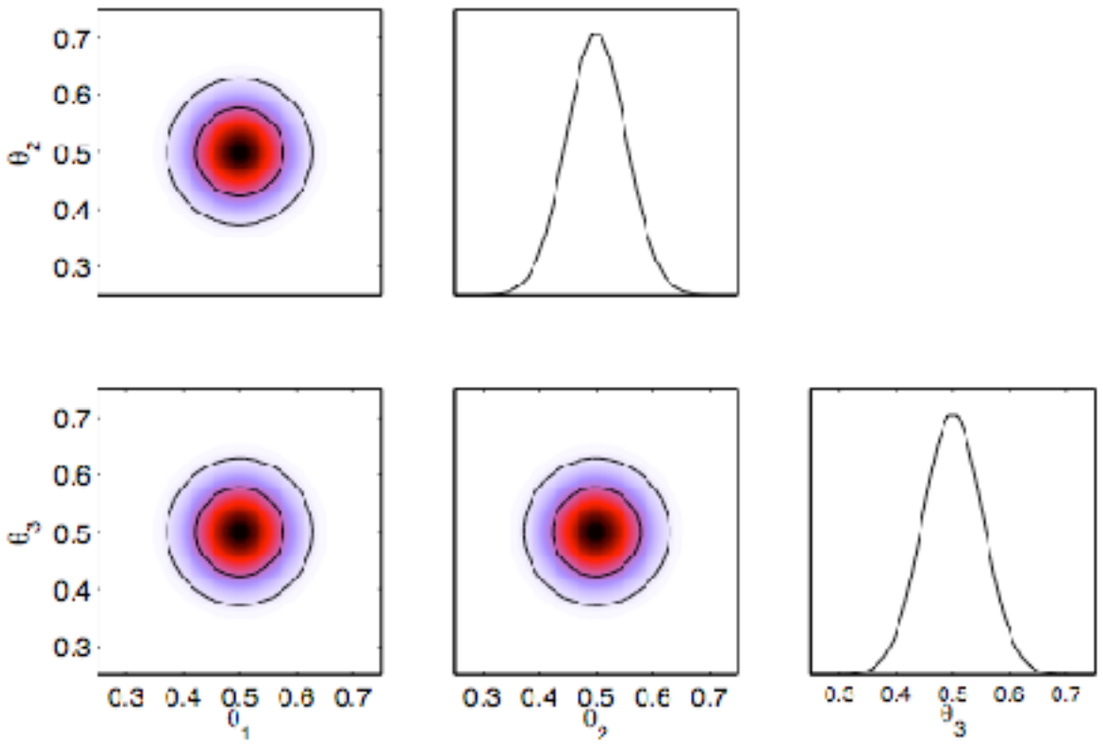
Marginalization fully propagates uncertainty onto the parameters of interest (verify: error propagation is a special case)

The Gaussian case

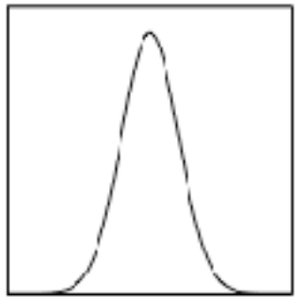
Imperial College London

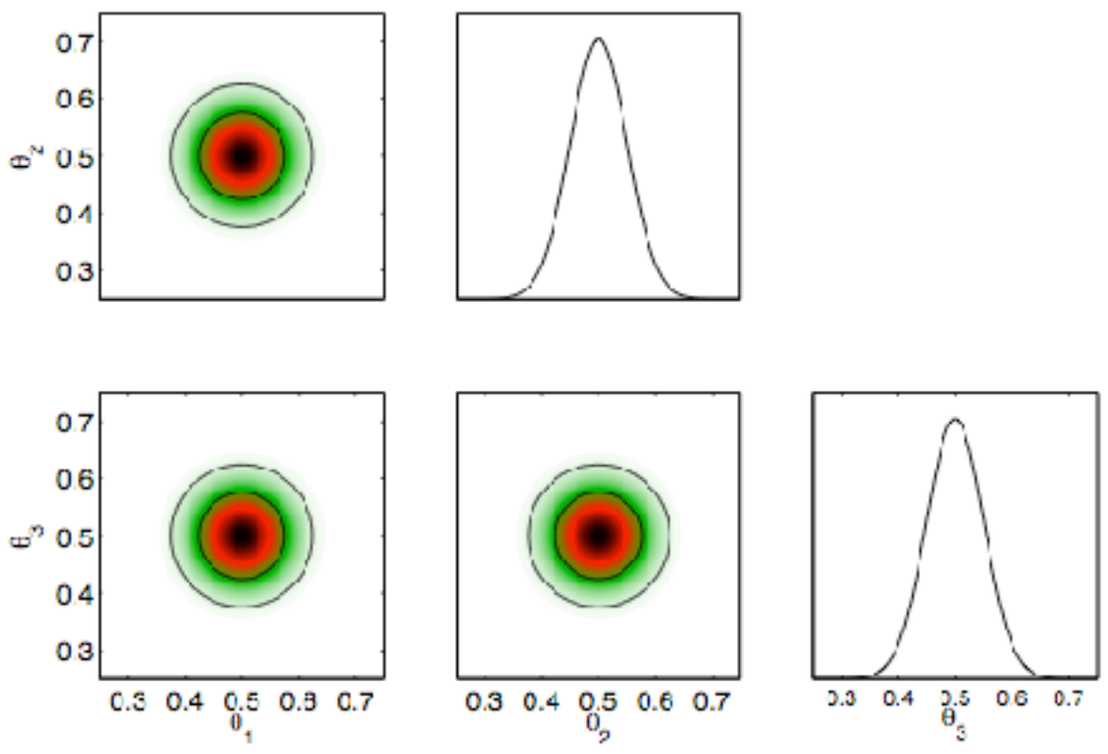
BAYESIAN: Marginal posterior





FREQUENTIST: Profile likelihood

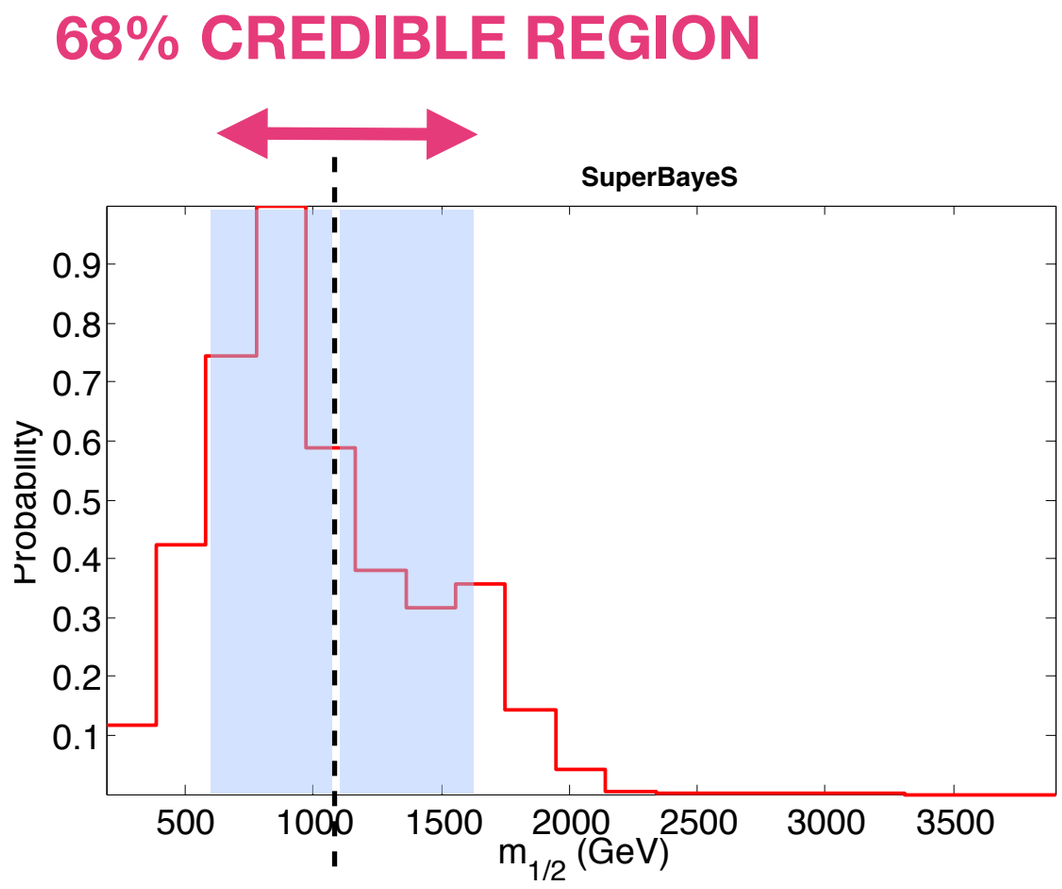




Credible regions: Bayesian approach

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)

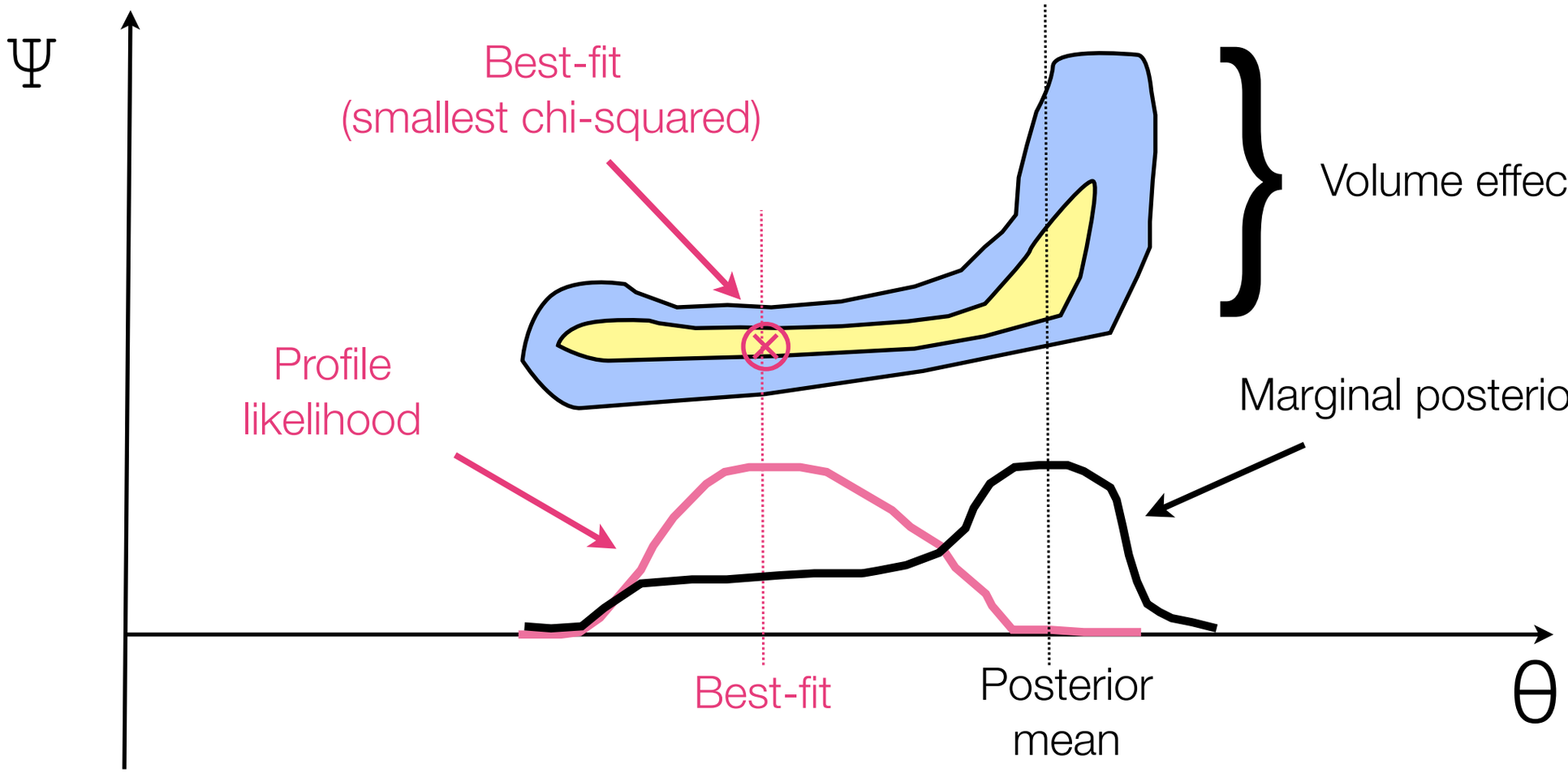
Imperial College London

Marginal posterior:

$$p(\theta|d) = \int d\psi p(\theta, \psi|d)$$

Profile likelihood:

$$\mathcal{L}(\theta) = \max_\psi \mathcal{L}(\theta, \psi)$$



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

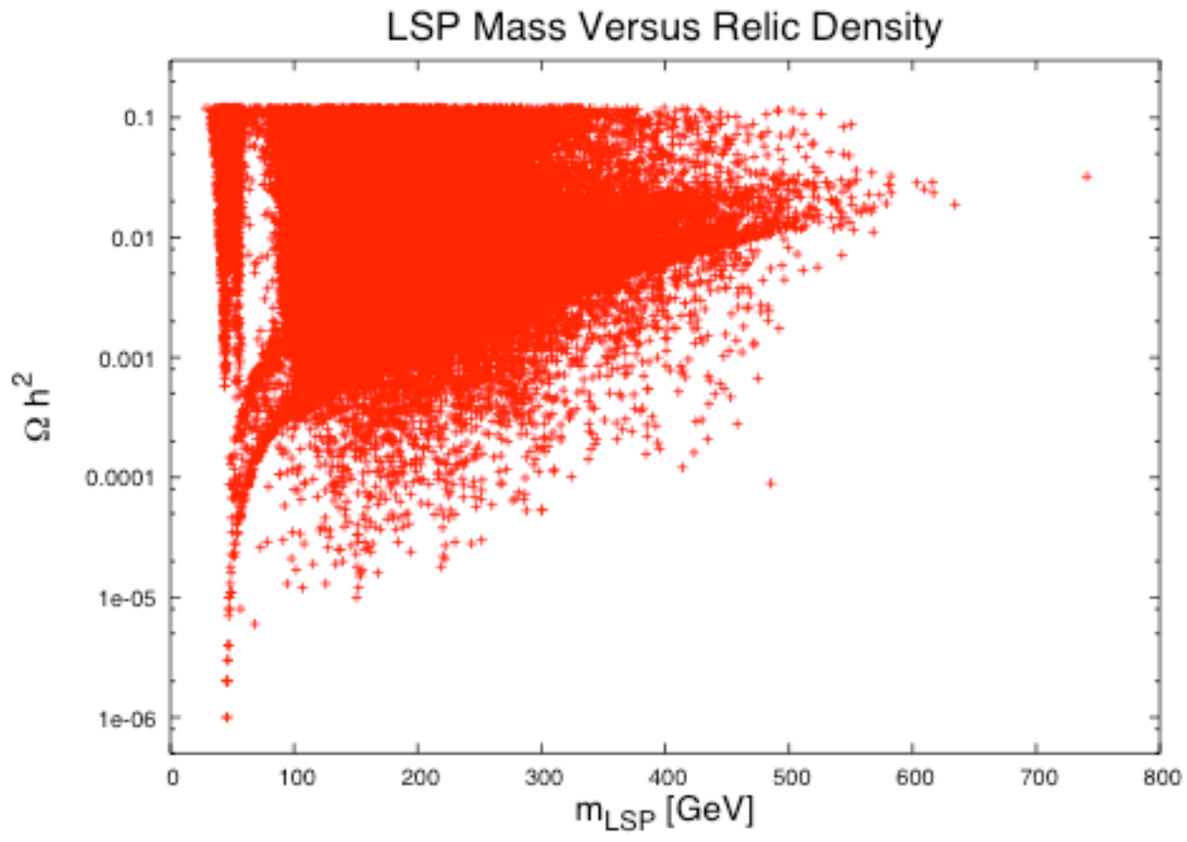
Markov Chain Monte Carlo

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!

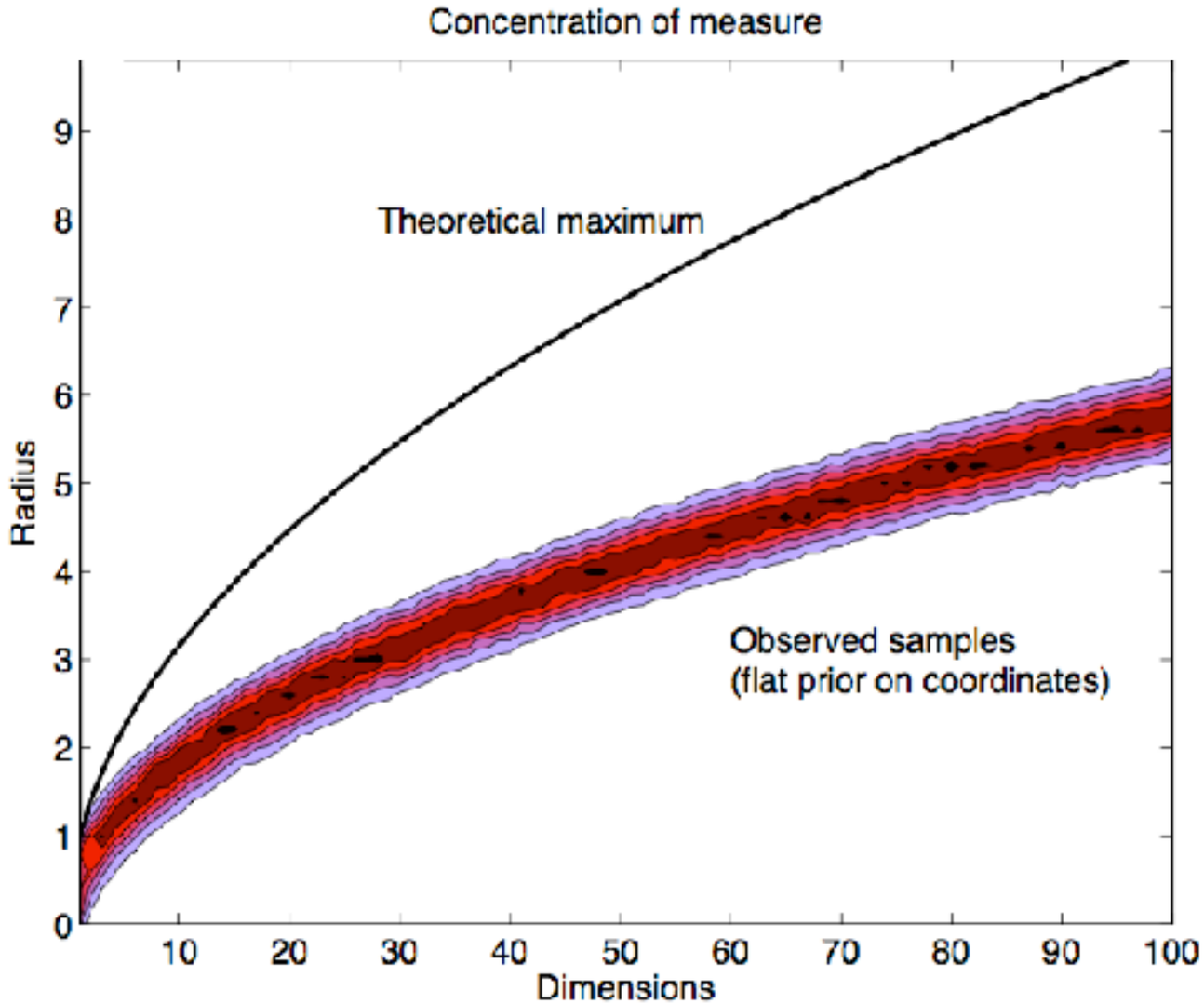
One recent example:
Berger et al (0812.0980)
pMSSM scans
(20 dimensions)

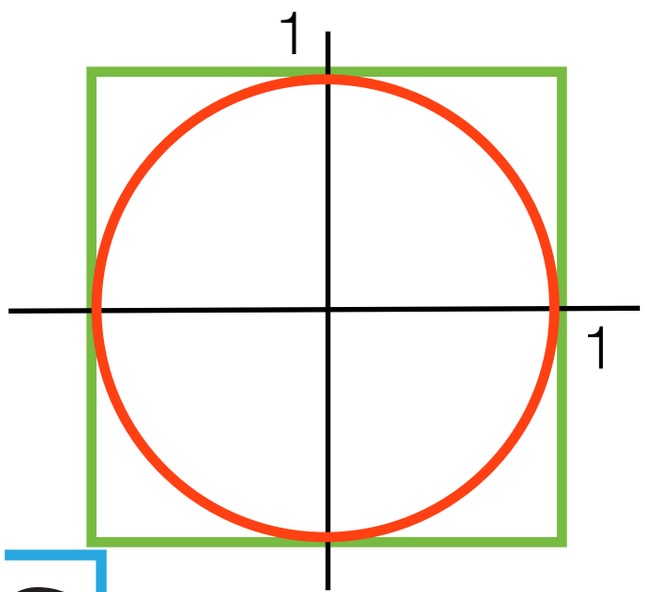


Random scans explore only a small fraction of the parameter space

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





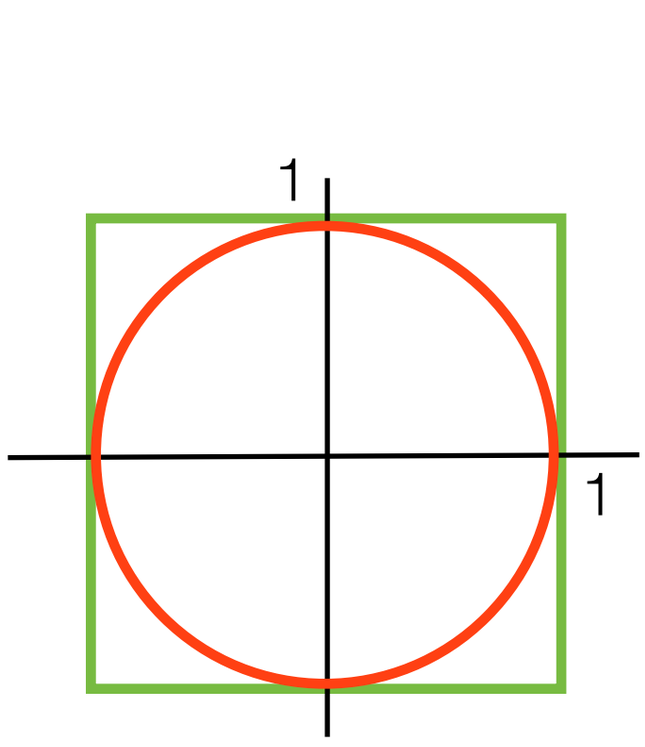
ICIC

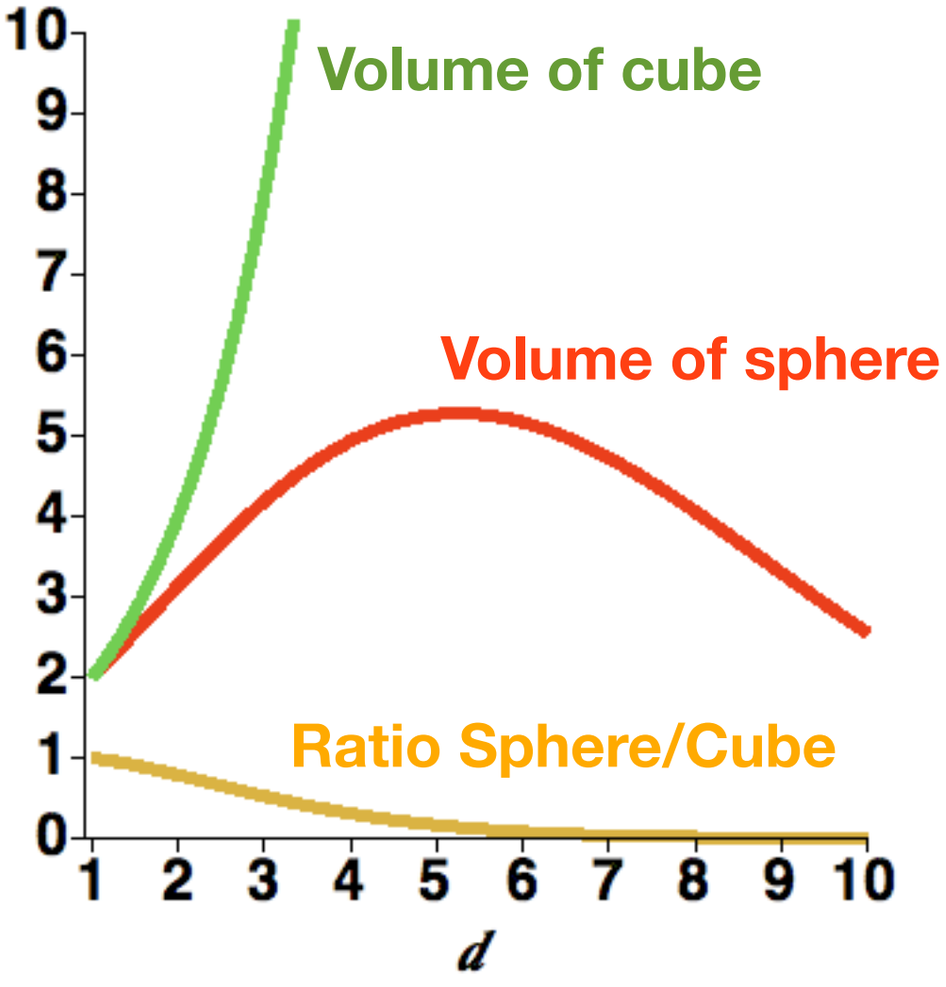
Geometry in high-D spaces

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.





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

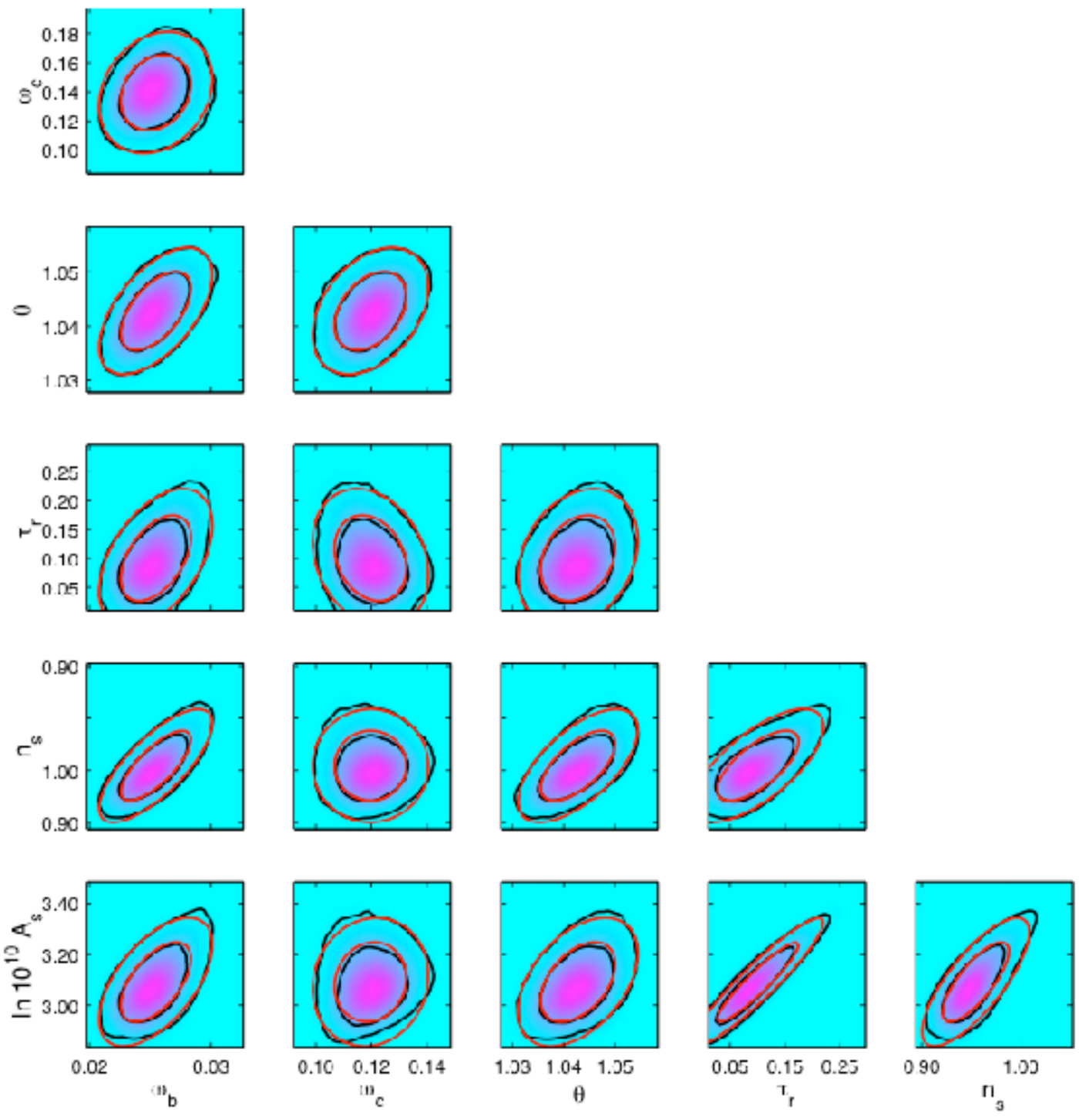
Imperial College London

- **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

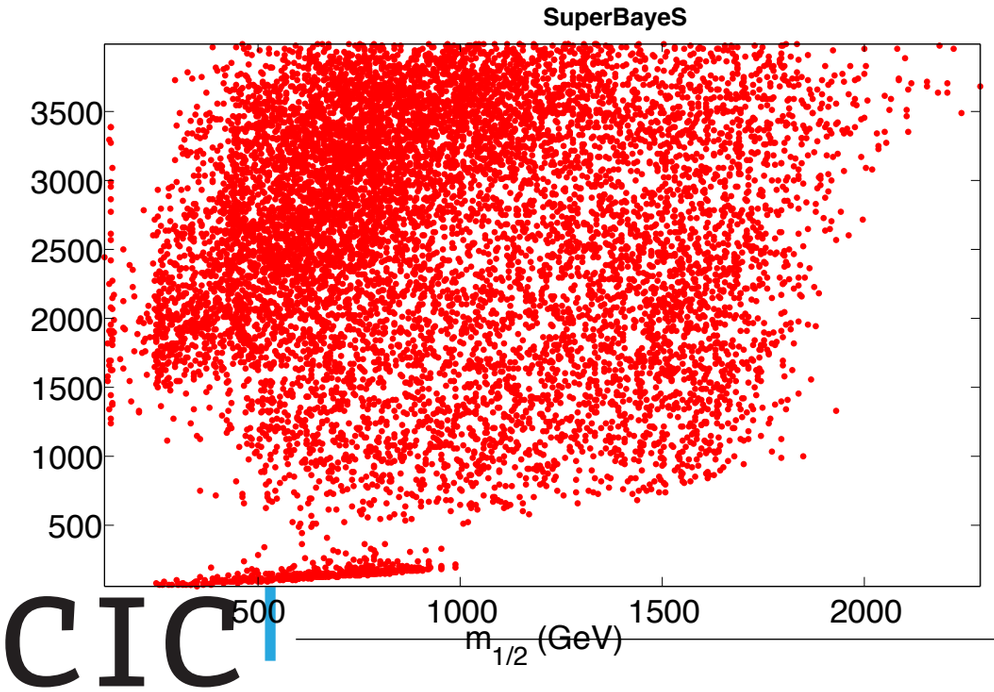


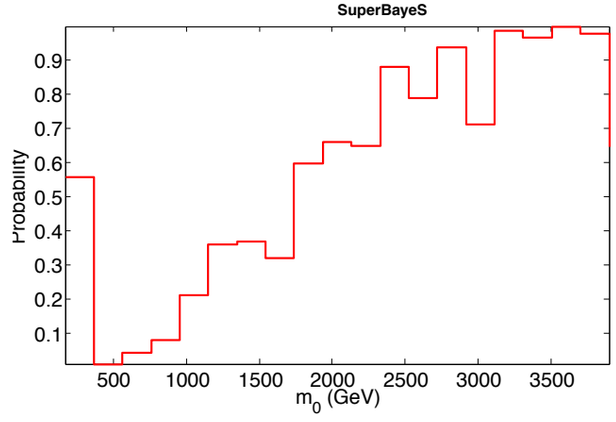
ICIC

MCMC estimation

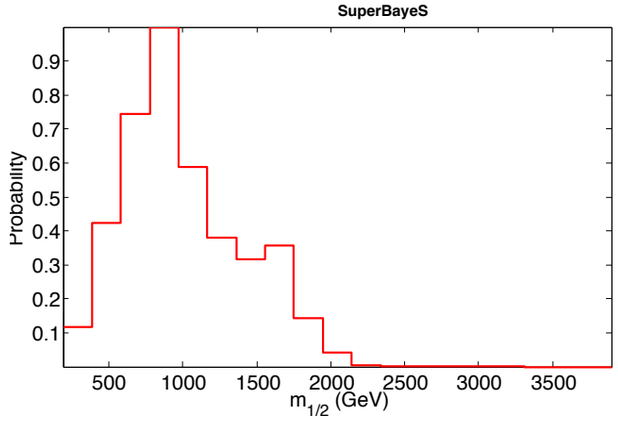
- **Marginalisation becomes trivial:** create bins along the dimension of interest and simply count samples falling within each bins ignoring all other coordinates
- Examples (from **superbayes.org**) :

2D distribution of samples from joint posterior





1D marginalised posterior (along y)

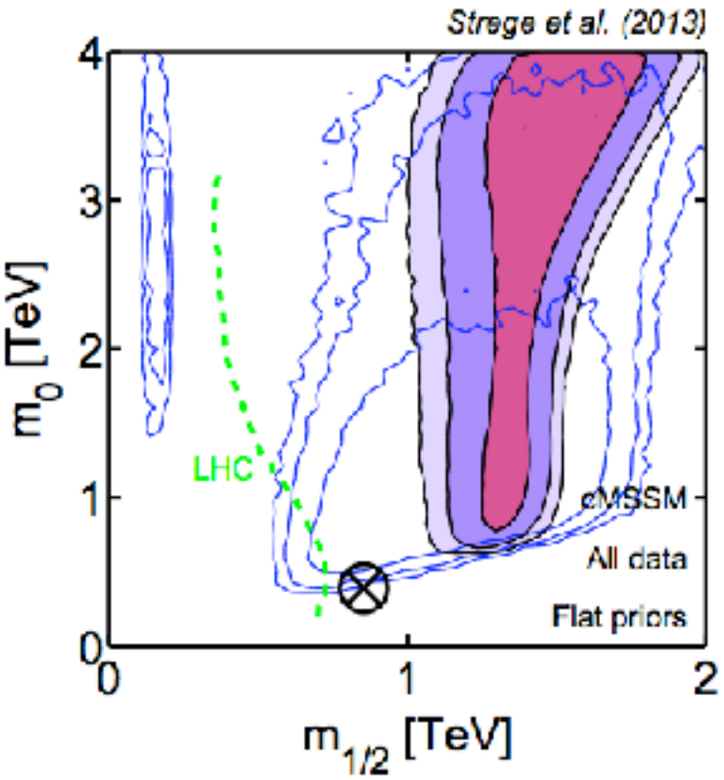


1D marginalised posterior (along x)

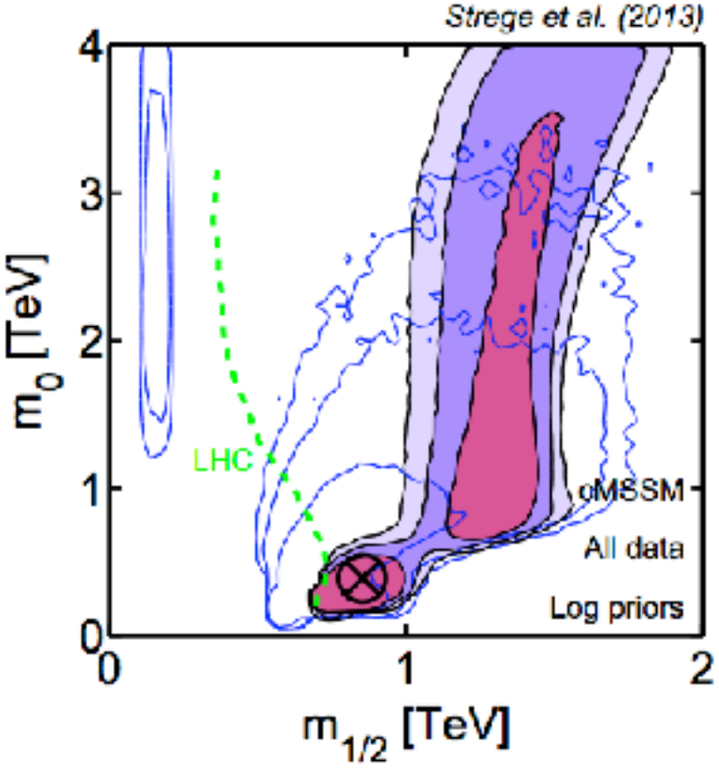
Non-Gaussian example

Imperial College London

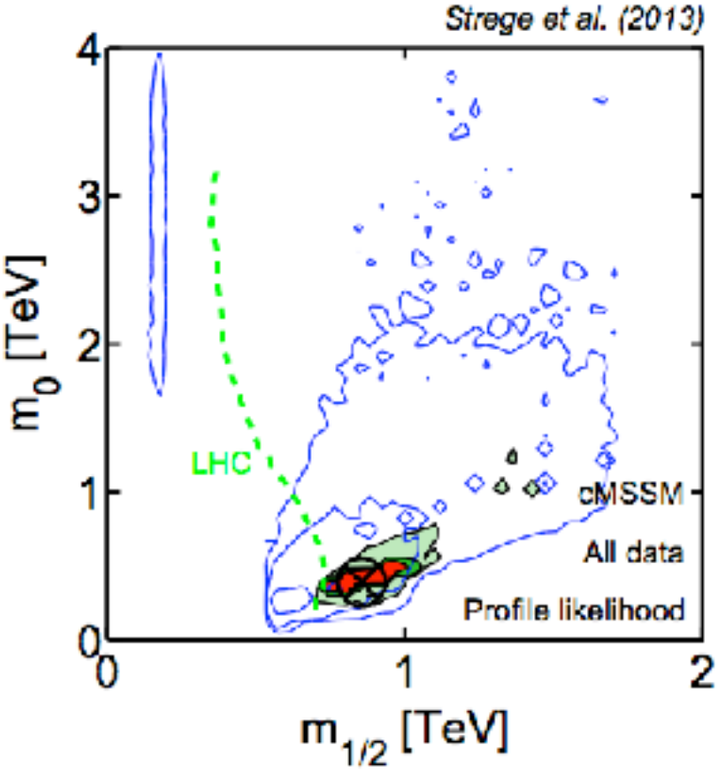
Bayesian posterior
("flat priors")



Bayesian posterior
("log priors")



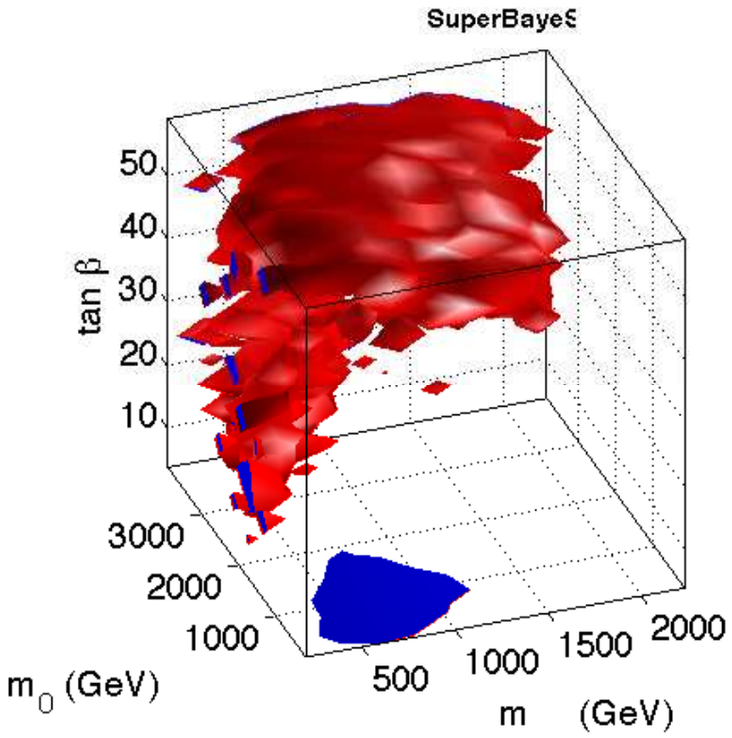
Profile likelihood

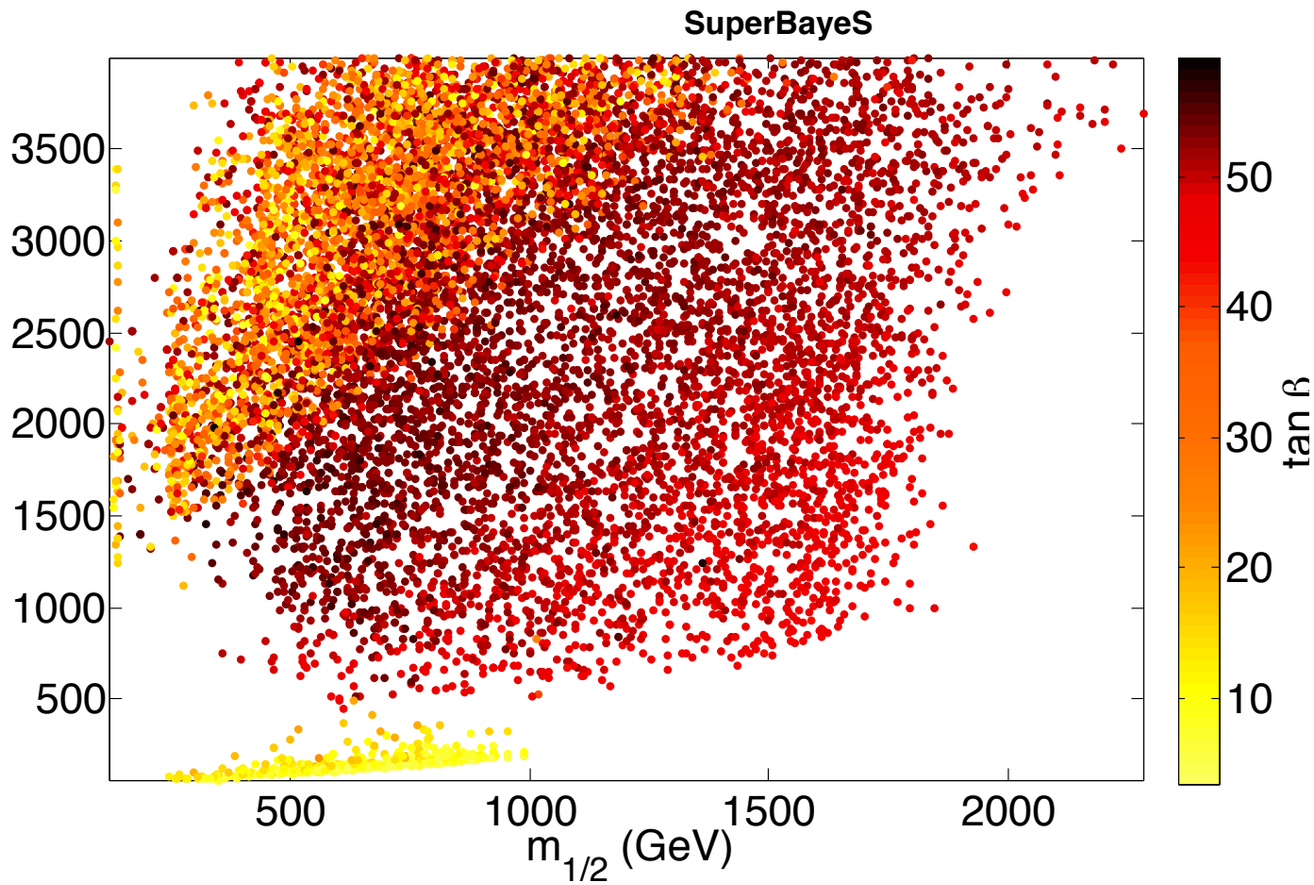


Constrained Minimal Supersymmetric Standard Model (4 parameters)
Strege, RT et al (2013)

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

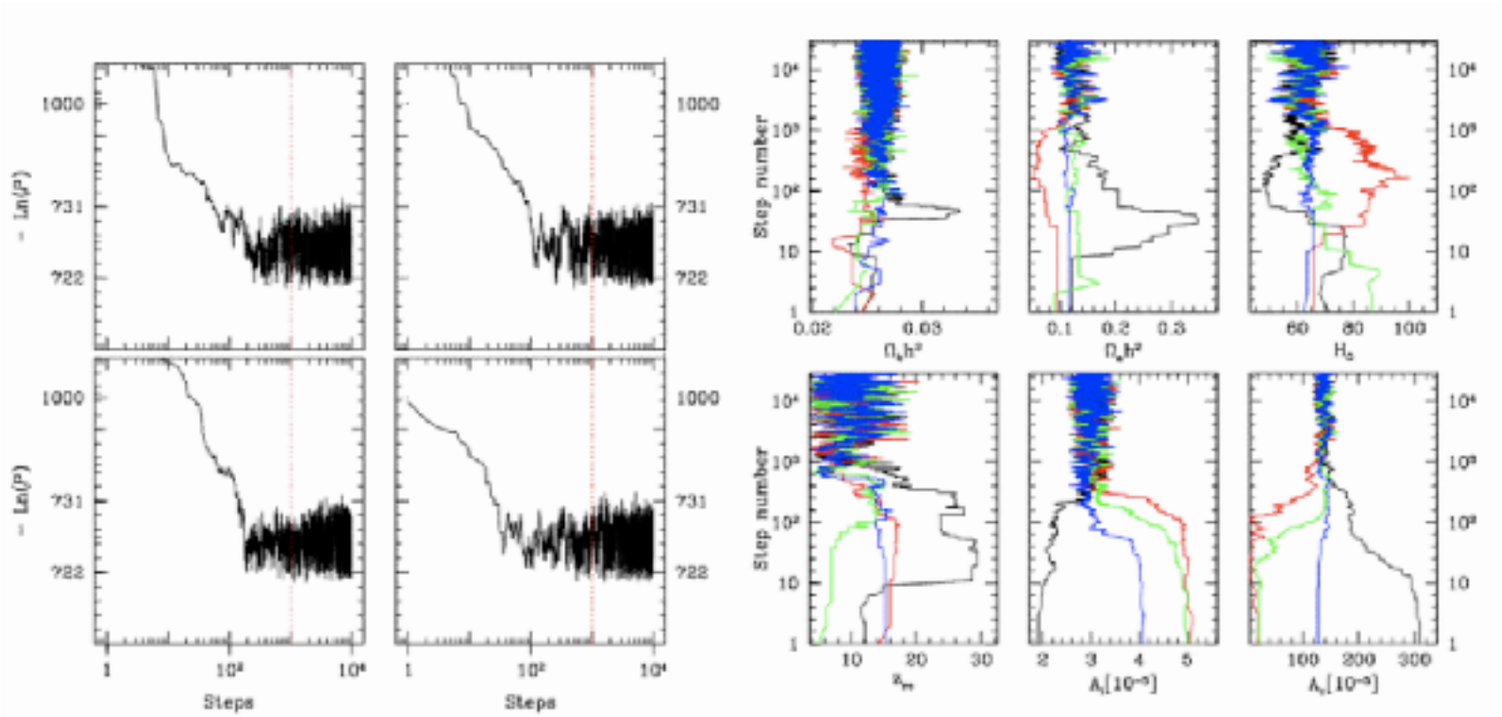
MCMC diagnostics

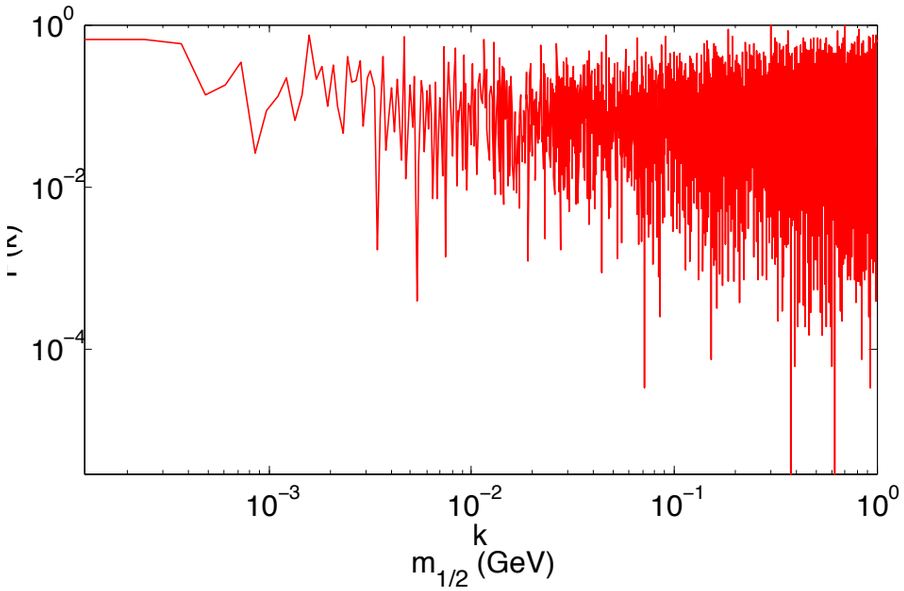
Imperial College London

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*)