Bayesian statistics

Third ASTERICS-OBELICS Workshop

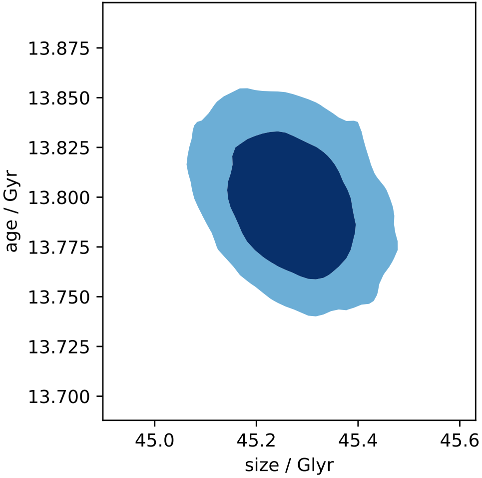
Will Handley
wh260@cam.ac.uk

Kavli Institute for Cosmology
Cavendish Laboratory (Astrophysics Group)
University of Cambridge

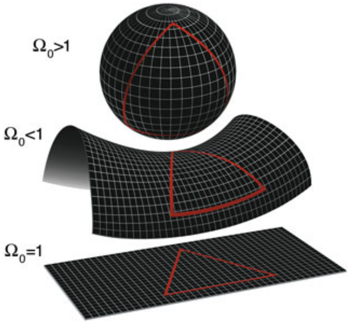
October 24, 2018

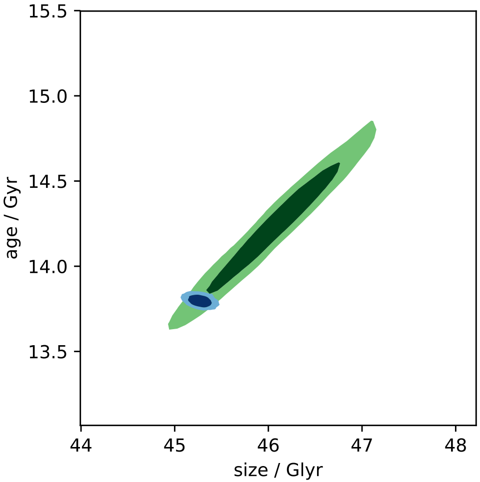
Inference in cosmology: parameter estimation

- Cosmologists infer universe parameters from data
- Bayesian framework: Use probability distributions to quantify errors
- Inferences depend on models (ΛCDM)
- arXiv:1807.06209



Inference in cosmology: model comparison





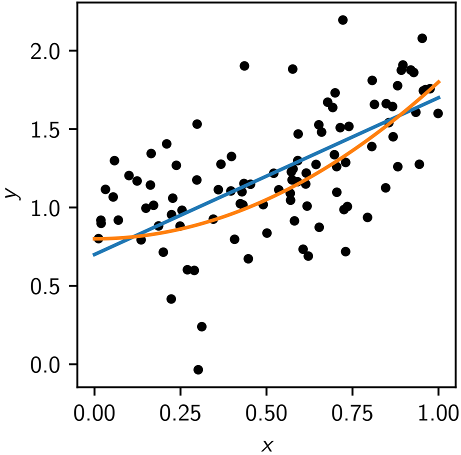
- Green model includes curvature (cΛCDM)
- Age and size now correlated
- Measurement less precise
- Flat is better with 2:1 odds against curvature

Motivating example: Fitting a line to data

- We have noisy data D
- We wish to fit a model M
- Functional form $y = f_M(x;\theta)$
- For example:

$$f_{\text{linear}}(x;\theta) = ax + b$$
$$f_{\text{quadratic}}(x;\theta) = ax^2 + b$$

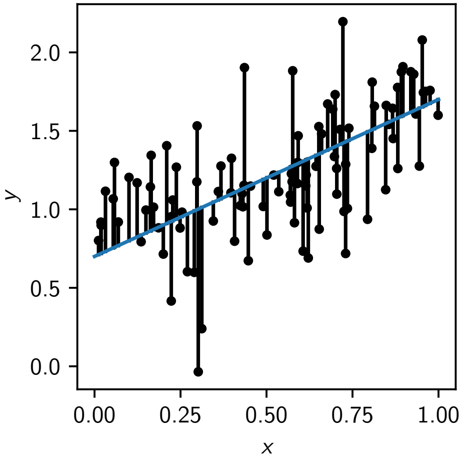
- Model parameters $\theta = (a, b)$



- For each parameter set θ:

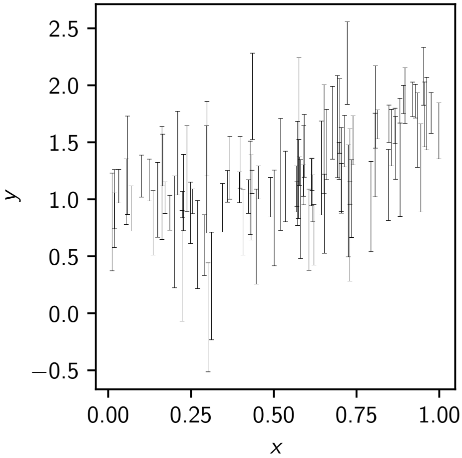
$$\chi^2(\theta) = \sum_i |y_i - f(x_i;\theta)|^2$$

- Minimise χ^2 wrt θ



- If data have non-uniform errors:

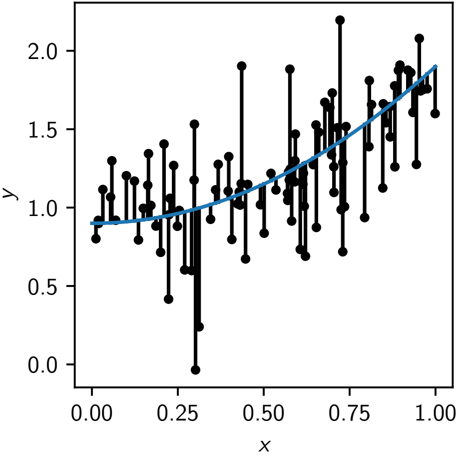
$$\chi^2(\theta) = \sum_i \frac{|y_i - f(x_i;\theta)|^2}{\sigma_i^2}$$



- Why square the errors? – could take absolute:

$$\psi^2(\theta) = \sum_i \frac{|y_i - f(x_i;\theta)|}{\sigma_i}$$

- How do we differentiate between models, e.g. quadratic vs curved

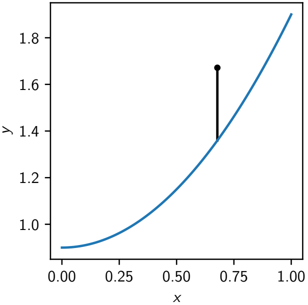


The probability of observing a datum:

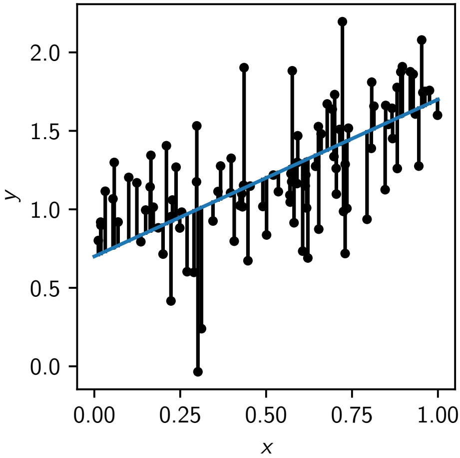
$$P(y_i|\theta, M) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{|y_i - f(x_i;\theta)|^2}{2\sigma_i^2}\right)$$

The probability of observing the data:

$$\begin{aligned} P(D|\theta, M) &= \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{|y_i - f(x_i;\theta)|^2}{2\sigma_i^2}\right) \\ &= \frac{1}{\prod_i \sqrt{2\pi}\sigma_i} \exp \sum_i -\frac{|y_i - f(x_i;\theta)|^2}{2\sigma_i^2} \\ &\propto e^{-\chi^2(\theta)/2} \end{aligned}$$



Maximum likelihood
Fitting lines to data



- Minimising $\chi^2(\theta)$ is equivalent to maximising $P(D|\theta, M) \propto e^{-\chi^2(\theta)/2}$
- $P(D|\theta, M)$ is called the Likelihood $L = L(\theta)$ of the parameters θ
- "Least squares" $\equiv$ "maximum likelihood" (if data are gaussian).
- arXiv:1809.04598

Bayesian inference

- Likelihood $L = P(D|\theta, M)$ is undeniably correct.
- Frequentists construct inference techniques purely from this function.
- The trend is cosmology is to work with a Bayesian approach.
- What we want are things like $P(\theta|D, M)$ and $P(M|D)$.
- To invert the conditionals, we need Bayes theorem:

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

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

Terminology

Bayesian inference

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

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

$$\text{Model probability} = \frac{\text{Evidence} \times \text{Model Prior}}{\text{Normalisation}}$$

The prior
Example: Biased coins

- Need to define the **Prior** $P(\theta)$ — probability of the bias, given no data
- Represents our knowledge of parameters before the data – subjective
- Frequentists view this as a flaw in Bayesian inference.
- Bayesians view this as an advantage
- Fundamental rule of Inference:

The prior

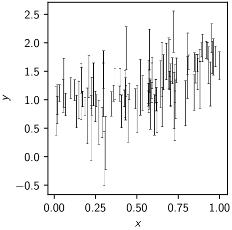
Example: Biased coins

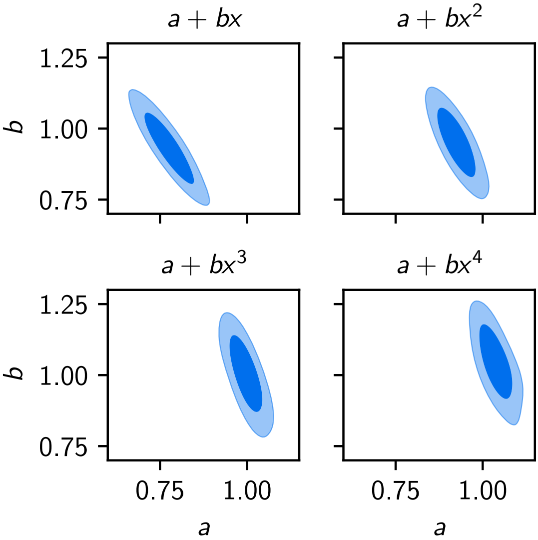
- Need to define the **Prior** $P(\theta)$ — probability of the bias, given no data
- Represents our knowledge of parameters before the data – subjective
- Frequentists view this as a flaw in Bayesian inference.
- Bayesians view this as an advantage
- Fundamental rule of Inference:

You cannot extract information from data
without making assumptions

- All Bayesians do is make them explicit
- Any method that claims it is "objective" is simply hiding them

- We may use $P(\theta|D, M)$ to inspect whether a model looks reasonable



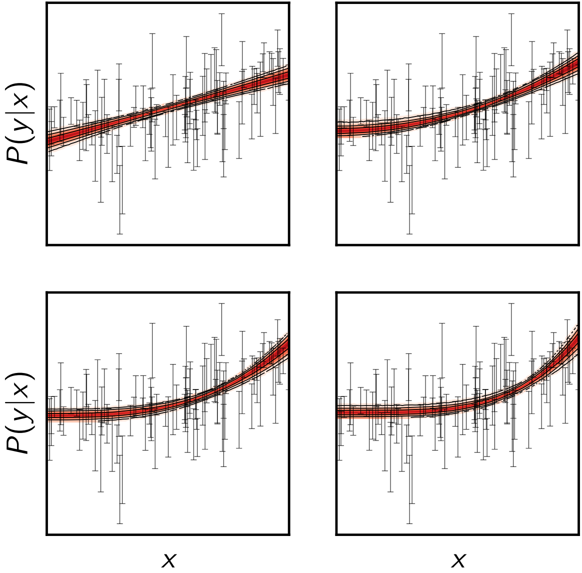


Predictive posterior

More useful to plot:

$P(y|x) = \int P(y|x,\theta)P(\theta)d\theta$

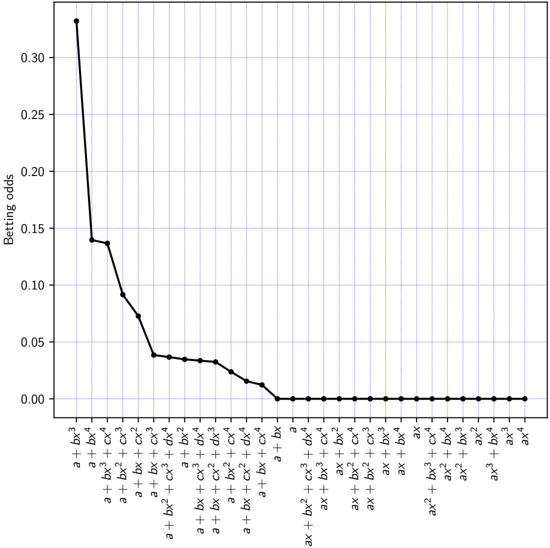
(all conditioned on D, M)



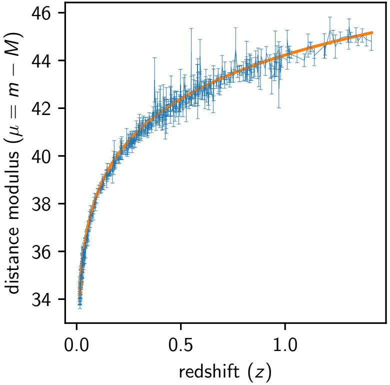
Model comparison

Bayesian inference

- We may use the Bayesian evidence Z to determine whether a model is reasonable.
- $Z = P(D|M) = \int P(D|M,\theta)P(\theta|M)d\theta$
- The evidence quantifies Occam's razor, penalising over-fitted models with too many parameters.
- Normally assume uniform model priors $Z \propto P(M|D)P(M)$.



- Whilst this model seems a little trite. . .
- . . . determining polynomial indices $\equiv$ determining cosmological material content:



$$\left(\frac{H}{H_0}\right)^2 = \Omega_\mathrm{r}\left(\frac{a_0}{a}\right)^4 + \Omega_\mathrm{m}\left(\frac{a_0}{a}\right)^3 + \Omega_k\left(\frac{a_0}{a}\right)^2 + \Omega_\Lambda$$

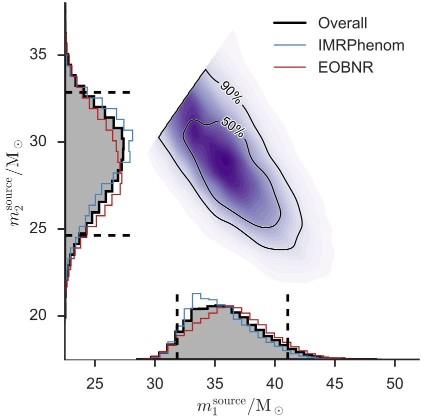
Quantifying error with Probability

- As scientists, we are used to seeing error bars on results.
- Masses of LIGO GW150914 binary merger:

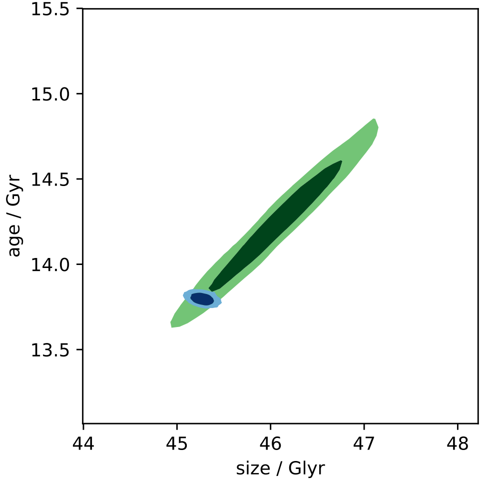
$$m_1 = 39.4^{+5.5}_{-4.9}\ M_\odot$$

$$m_2 = 30.9^{+4.8}_{-4.4}\ M_\odot$$

- These are called *credible intervals*, state that we are e.g. 90% confident of the value lying in this range.
- More importantly, these are *summary statistics*.



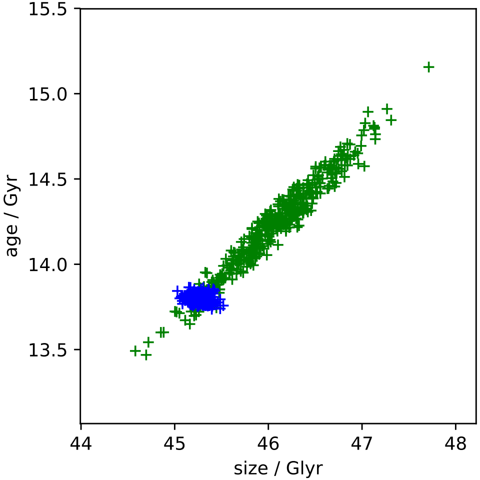
- In high dimensions, posterior $\mathcal{P}$ occupies a vanishingly small region of the prior π.
- Gridding is doomed to failure for $D \gtrsim 4$.
- *Sampling* the posterior is an excellent compression scheme.
- Name of the game: Constructing algorithms to generate samples with a minimum number of likelihood calls



Sampling

How to describe a high-dimensional posterior

- In high dimensions, posterior $\mathcal{P}$ occupies a vanishingly small region of the prior π.
- Gridding is doomed to failure for $D \gtrsim 4$.
- *Sampling* the posterior is an excellent compression scheme.
- Name of the game: Constructing algorithms to generate samples with a minimum number of likelihood calls



Sampling algorithms: Metropolis Hastings

- Turn the N-dimensional problem into a one-dimensional one.
  1. Propose random step to new point $x_i \to x_{i+1}$
  2. If uphill $[P(x_{i+1}) > P(x_i)]$, make step...
  3. ...otherwise make step with probability $\propto P(x_{i+1})/P(x_i)$.
- Theorem: set of steps $\{x_i : i = 1 \ldots N\}$ are samples from posterior P
- chi-feng.github.io/mcmc-demo/app.html#RandomWalkMH,banana

Hamiltonian Monte-Carlo

- Key idea: Treat $\log L(\Theta)$ as a potential energy
- Guide walker under force:

$$F(\Theta) = \nabla \log L(\Theta)$$

- Walker is naturally guided uphill
- Conserved quantities mean efficient acceptance ratios.
- Allows sampling in millions of dimensions.
- stan is a fully fledged probabilistic programming language for HMC (10.18637/jss.v076.i01).
- chi-feng.github.io/mcmc-demo/app.html#HamiltonianMC,donut

Ensemble sampling

- Instead of one walker, evolve a set of n walkers.
- Can use information present in ensemble to guide proposals.
- emcee: affine invariant proposals arXiv:1202.3665
- chi-feng.github.io/mcmc-demo/app.html#SVGD,banana

Nested Sampling

John Skilling's alternative to traditional MCMC

- Uses ensemble sampling to compress prior to posterior.
- Allows you to compute evidences, partition functions and Kullback-Liebler divergences.

New procedure:
Maintain a set S of n samples, which are sequentially updated:

S_0: Generate n samples uniformly over the space .

S_{n+1}: Delete the lowest probability sample in S_n, and replace it with a new sample with higher probability

Requires one to be able to uniformly within a region, subject to a *hard probability constraint*.

MultiNest Rejection sampling $D < 20$ (arXiv:0809.3437)

PolyChord Slice sampling $D \lesssim 1000$ (arXiv:1506.00171)

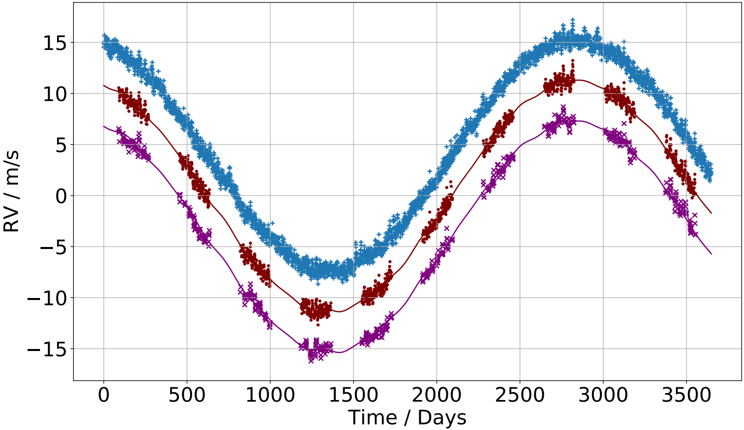
Sampling algorithms: summary

Metropolis Hastings	Easy to implement, requires manual tuning & insight into the problem
emcee	Fire-and-forget, easy python implementation
Hamiltonian Monte Carlo	Allows sampling in extremely high dimensions, requires gradients, self-tuning. Need to learn stan programming language.
Nested Sampling	Allows evidence calculation in moderately high dimensions, self-tuning. Need to install MultiNest and/or PolyChord packages.

Further Reading

- Data Analysis: A Bayesian Tutorial (Sivia & Skilling)
- Information theory, inference & learning algorithms (MacKay)
- Bayesian methods in cosmology arXiv.org:0803.4089
- Bayesian sparse reconstruction arXiv:1809.04598
- Hamiltonian monte carlo arXiv:1701.02434
- Nested sampling euclid.ba/1340370944

Example: Exoplanets



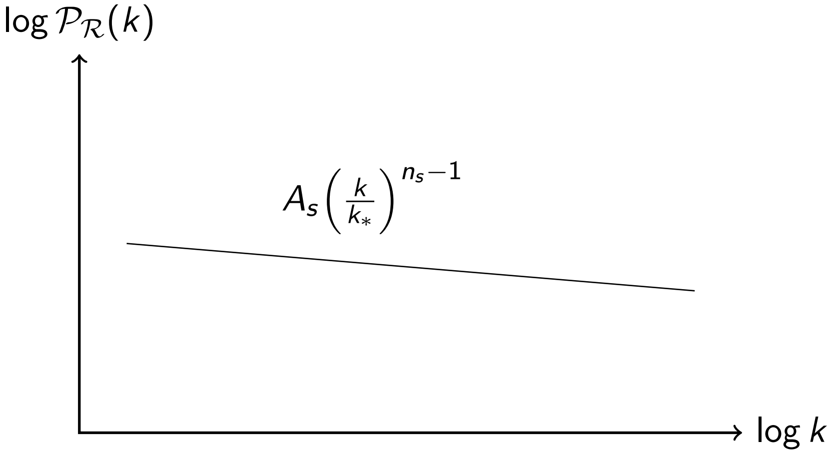
Example: Exoplanets

- Simple radial velocity model

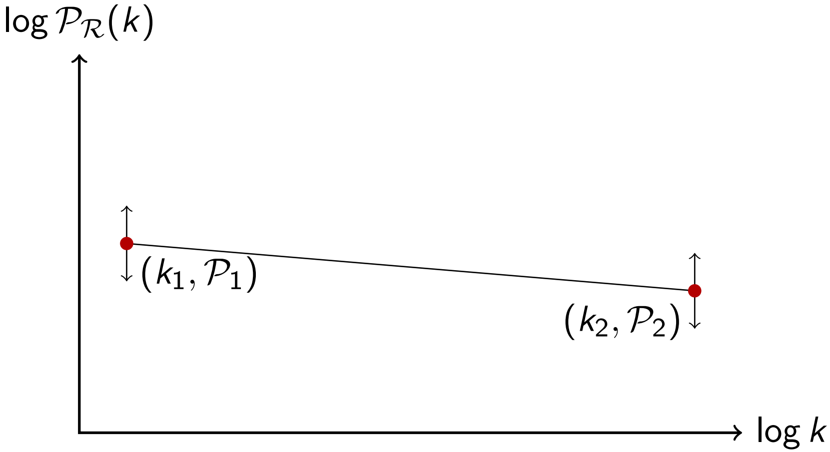
$$\nu(t;\theta) = \sum_{p=1}^{N} K_p \sin(\omega_p t + \phi_p)$$

- Fit each model to data.
- Posteriors on model parameters $[(K_p, \omega_p, \phi_p), p = 1 \cdots N]$ quantify knowledge of system characteristics.
- Evidences of models determine relative likelihood of number of planets in system
- arXiv:1806.00518

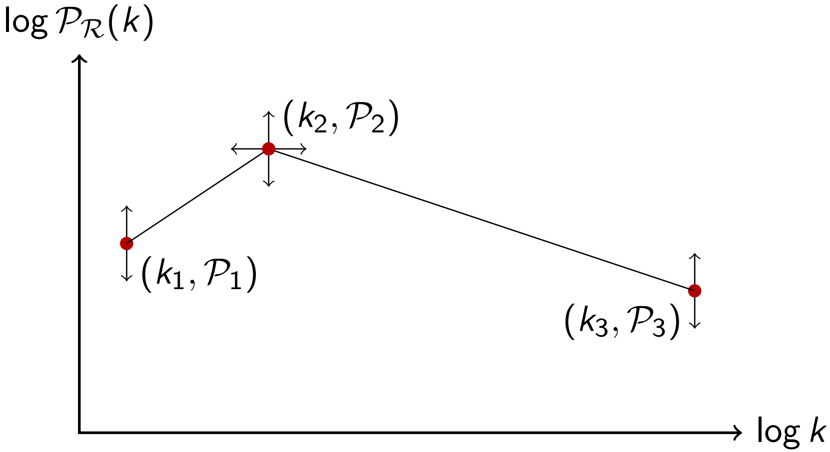
Example: function $\mathcal{P}_\mathcal{R}(k)$ reconstruction



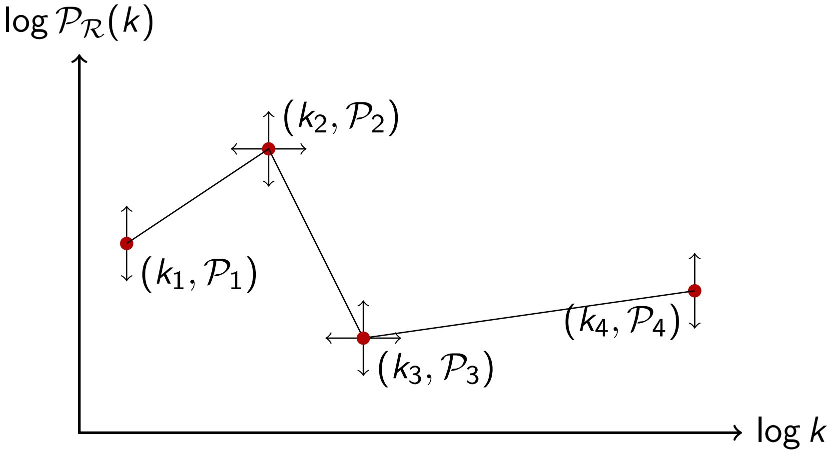
Example: function $\mathcal{P}_\mathcal{R}(k)$ reconstruction



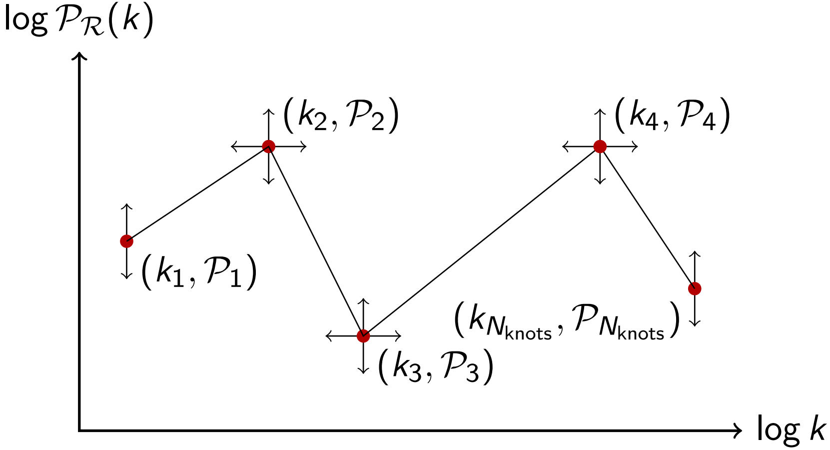
Example: function $\mathcal{P}_\mathcal{R}(k)$ reconstruction



Example: function $\mathcal{P}_\mathcal{R}(k)$ reconstruction

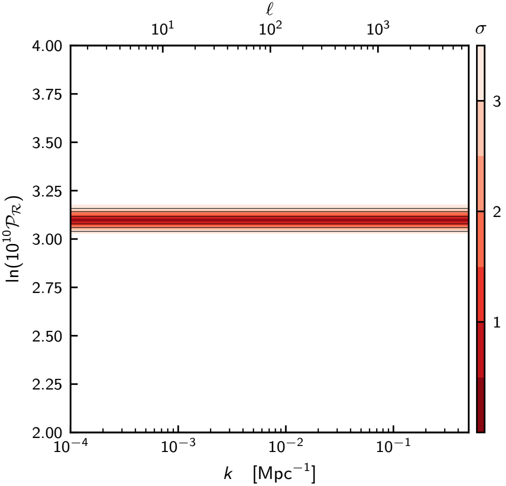


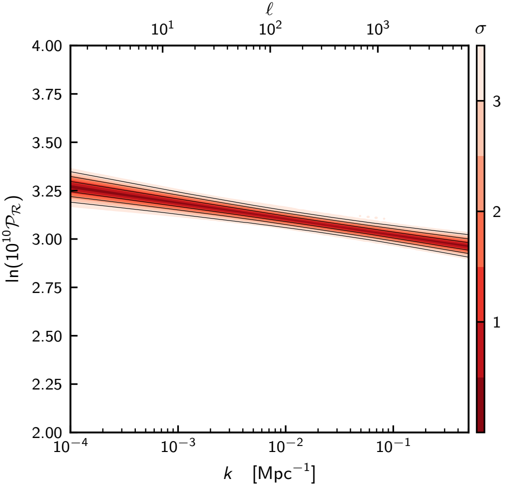
Example: function $\mathcal{P}_\mathcal{R}(k)$ reconstruction



no tilt

Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction

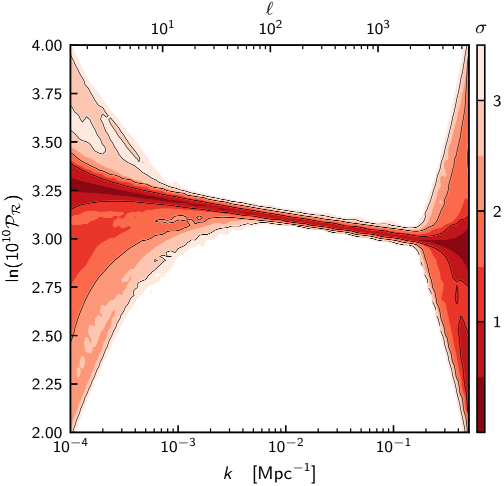




ℓ
10^1
10^2
10^3
σ
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
$\ln(10^{10}\mathcal{P}_\mathcal{R})$
3
2
1
10^{-4}
10^{-3}
10^{-2}
10^{-1}
$k \quad [\mathrm{Mpc}^{-1}]$

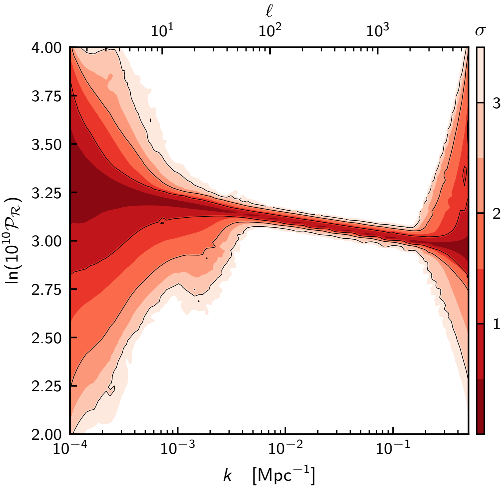
1 internal knot

Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction



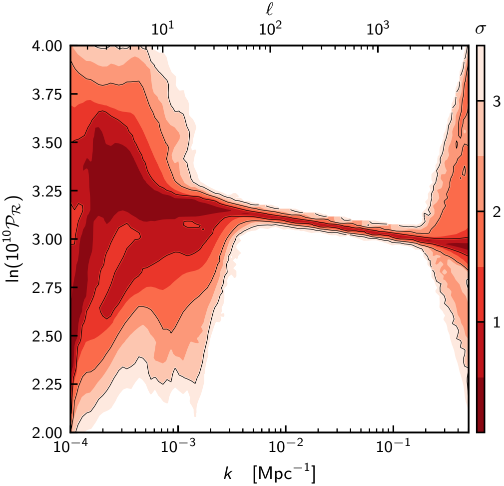
2 internal knots

Primordial power spectrum $\mathcal{P}_{\mathcal{R}}(k)$ reconstruction



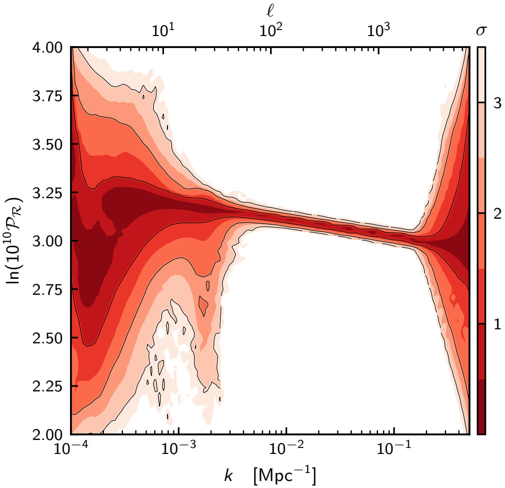
3 internal knots

Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction



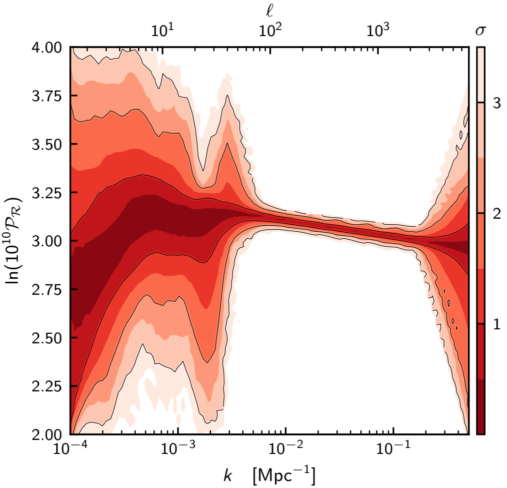
4 internal knots

Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction



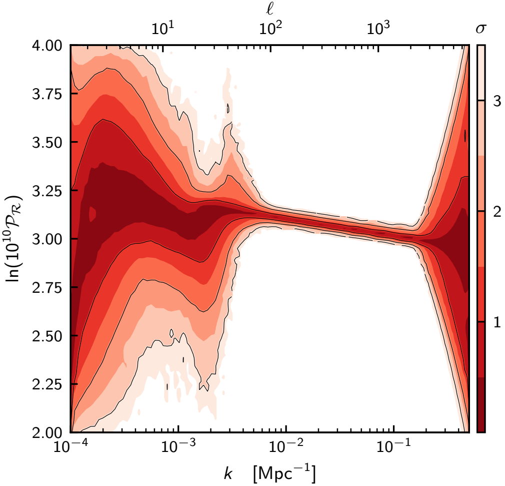
5 internal knots

Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction



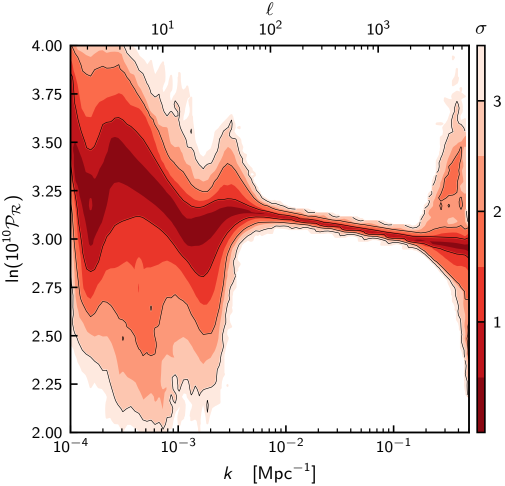
6 internal knots

Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction



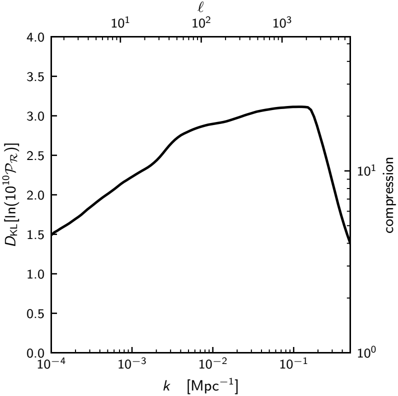
7 internal knots

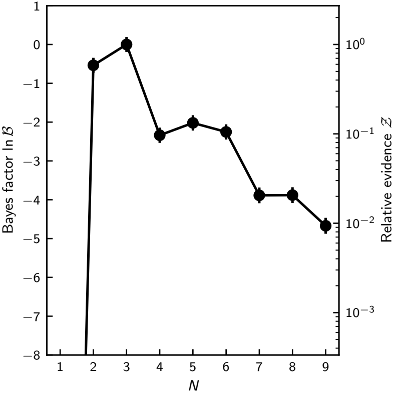
Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction



Bayes Factors

Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction





Marginalised plot

Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction