Statistics

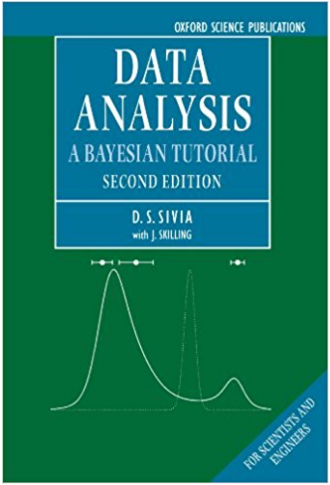
Aachen Cosmotools 2018

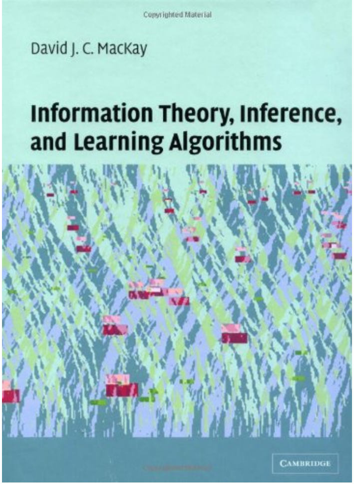
Will Handley
wh260@cam.ac.uk

Astrophysics Group
Cavendish Laboratory
University of Cambridge

April 24, 2018

Further reading





- Data analysis: A Bayesian Tutorial (Sivia & Skilling)
- Information Theory, Inference and Learning Algorithms (Mackay)

Introduction

- Statistics $\equiv$ Inference $\equiv$ Machine Learning/AI.
- How to extract information about scientific models from data.
- Most cosmologists work in a *Bayesian* framework of inference, although *Frequentist* methods are also sometimes used.

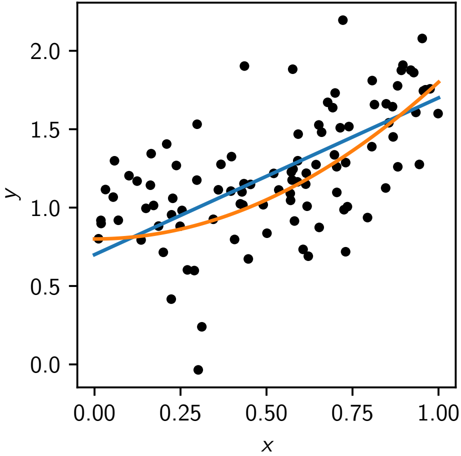
Motivating example

Fitting lines to data

- We have noisy data D
- We wish to fit a model M
- Functional form
  $y = f_M(x;\theta) = ax + b$
- e.g:

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

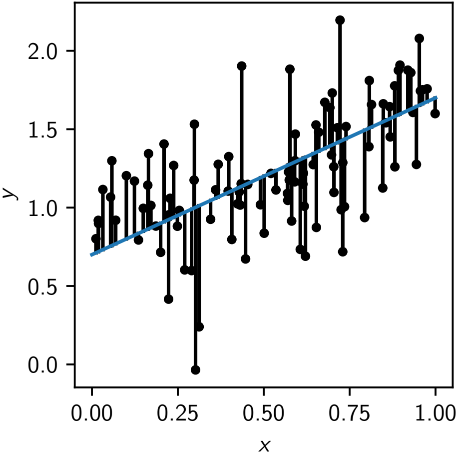
- Model parameters
  $\theta = (a, b)$



- For each parameter set θ:

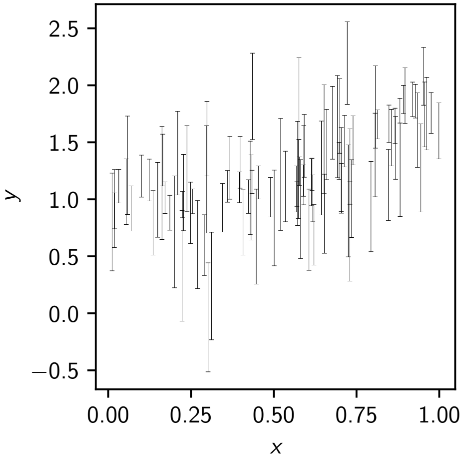
$$\chi^2(\theta) = \sum_i |y_i - f(x_i;\theta)|^2$$

- Minimise χ^2 wrt θ



- If data have non-uniform errors:

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



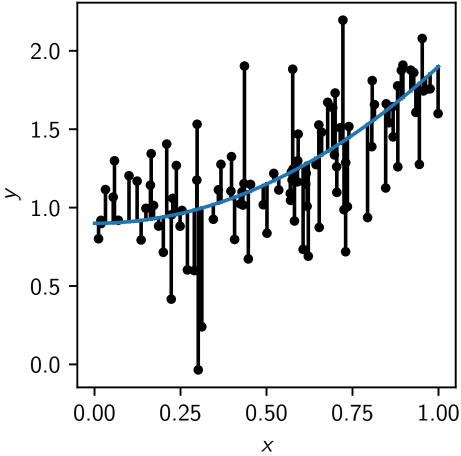
Problems with χ^2

Fitting lines to data

- How do we differentiate between models
- Why square the errors? – could take absolute:

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

- Where does this even come from?



Multivariate probability

- Marginalisation:
$$P(x) = \int P(x,y)dy$$
- Conditioning:
$$P(y|x) = \frac{P(x,y)}{P(x)} = \frac{P(x,y)}{\int P(x,y)dy}$$
- De-Conditioning:
$$P(x|y)P(y) = P(x,y)$$
- Bayes theorem:
$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

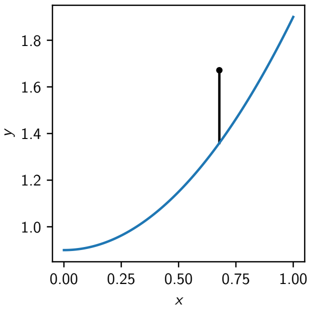
"To flip a conditional $P(x|y)$, you first de-condition on y, and then re-condition on x."

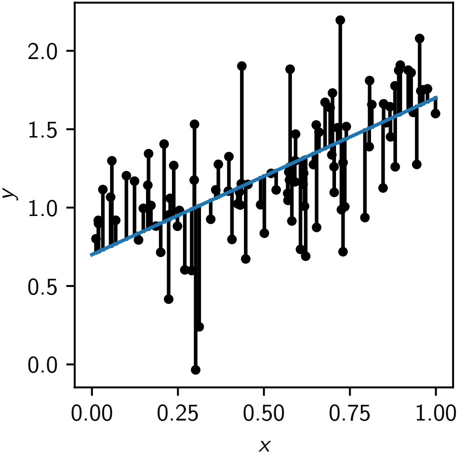
The probability of observing a datum:

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

The probability of observing the data:

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





- 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).

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

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

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

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

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

wh260@cam.ac.uk

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

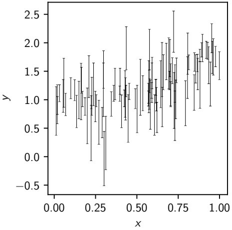
Example: Biased coins

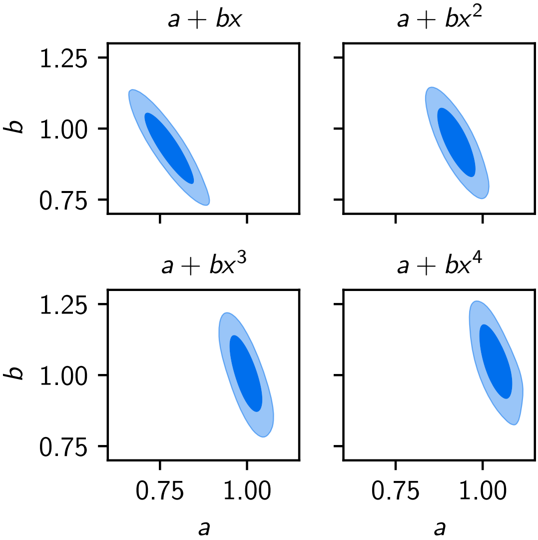
- Need to define the **Prior** $P(\theta)$ — probability of the bias, given no data
- Represents our knowledge of parameters before the data – subjective
- Frequentists view this as a flaw in Bayesian inference.
- Bayesians view this as an advantage
- Fundamental rule of Inference:

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

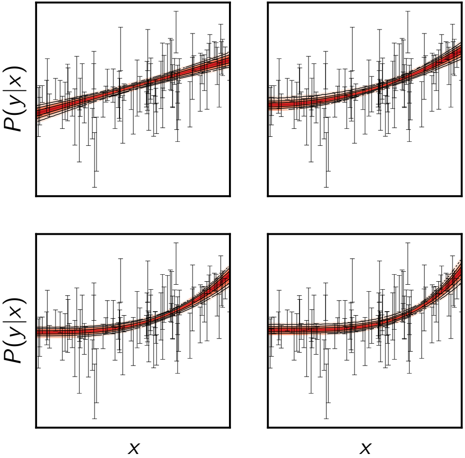




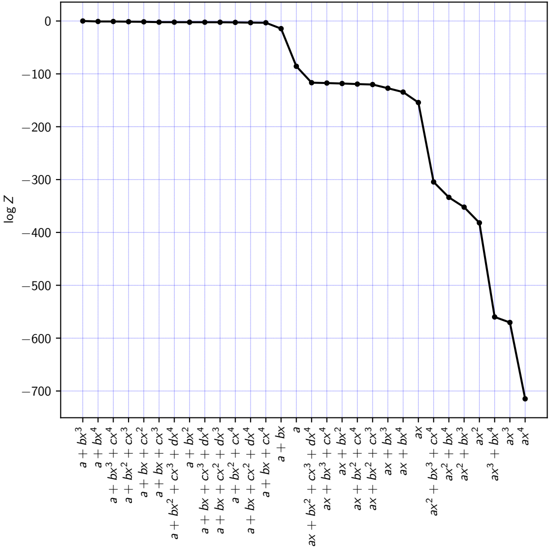
More useful to plot:

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

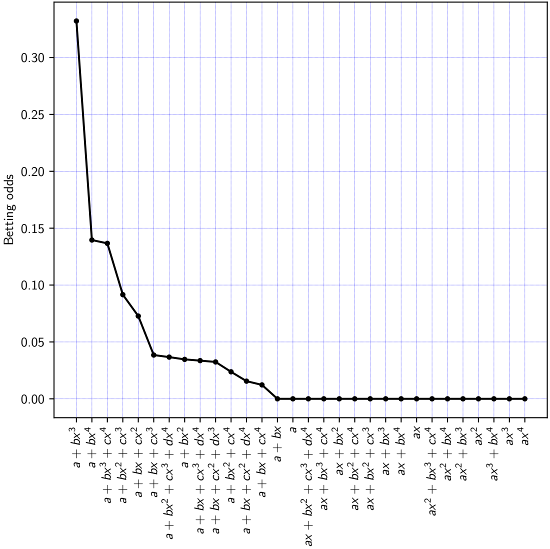
(all conditioned on D, M)



- We may use $Z = P(D|M)$ to determine whether a model is reasonable.

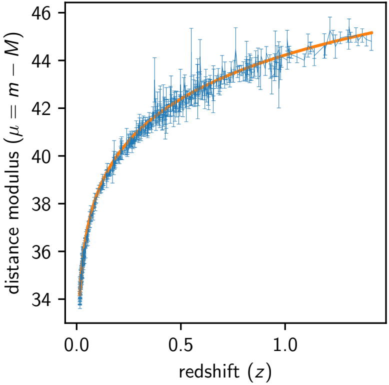


- We may use $Z = P(D|M)$ to determine whether a model is reasonable.



Line fitting (context)

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



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

Probability distributions

- As scientists, we are used to seeing error bars on results.
- Age of the universe (*Planck*):

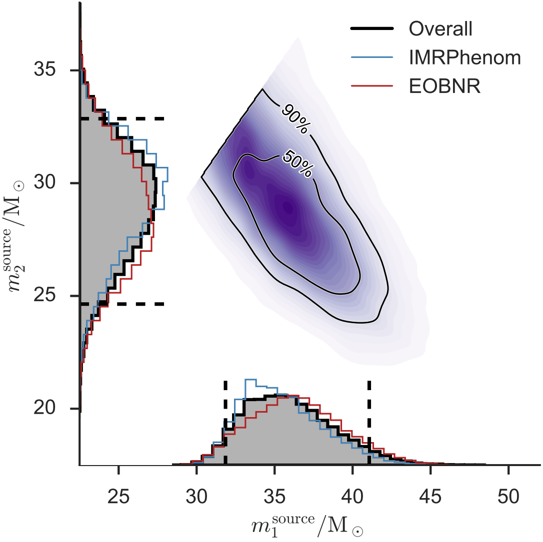
$$13.73 \pm 0.12 \,\text{billion years old.}$$

- Masses of LIGO GW150914 binary merger:

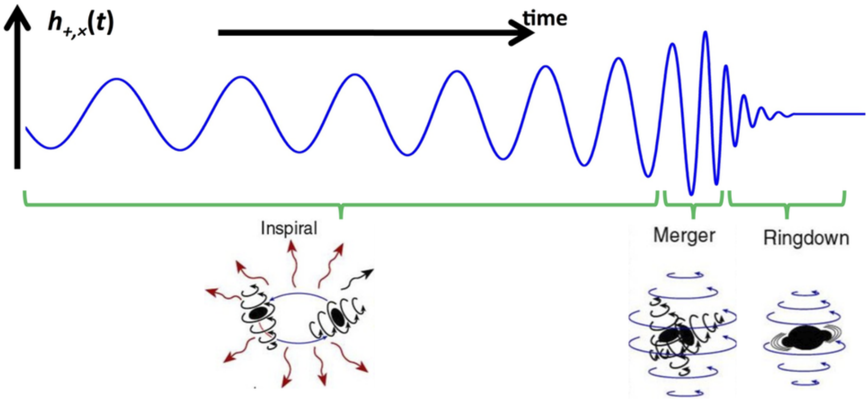
$$m_1 = 39.4^{+5.5}_{-4.9}\,M_\odot, \qquad m_2 = 30.9^{+4.8}_{-4.4}\,M_\odot$$

- 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*.

LIGO binary merger



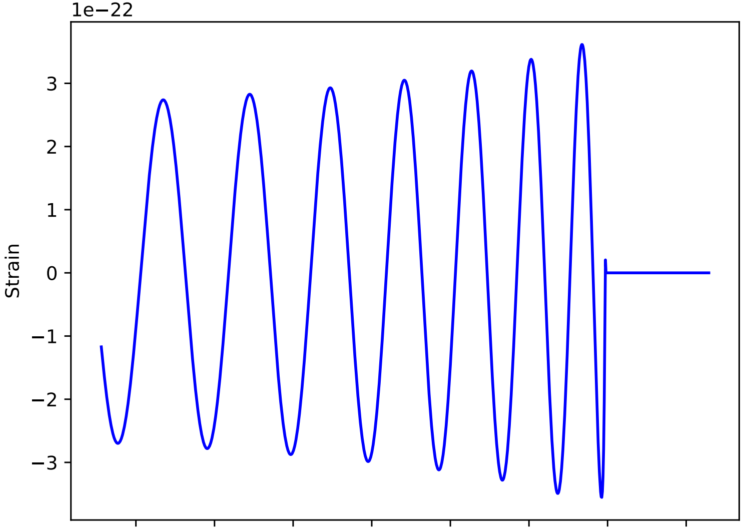
- Summary statistics summarise a full probability distribution.
- One goal of inference is to produce these probability distributions.



$h_{+,\times}(t)$
time
Inspiral
Merger
Ringdown

The model M

Extended example of inference: LIGO



The parameters Θ of the model M

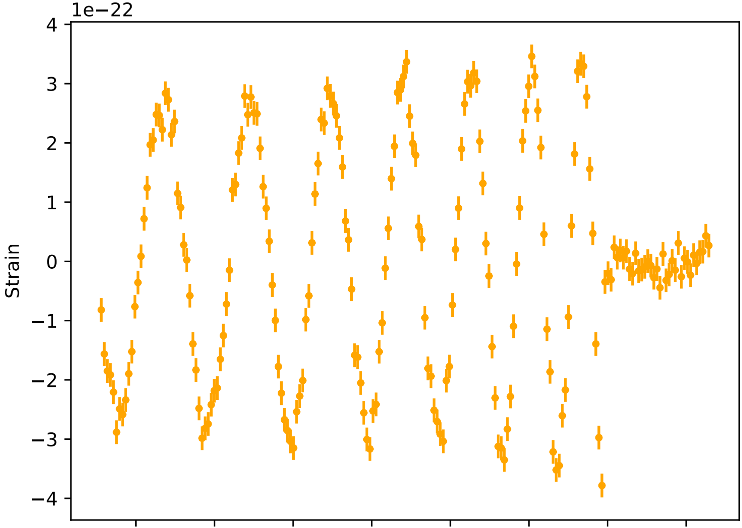
Extended example of inference: LIGO

Theoretical signal depends on:

- m_1, m_2: mass of binary
- θ, ϕ: sky location
- r: luminosity distance
- Φ_c, t_c: phase and time of coalescence
- i, θ_{sky}: inclination and angle on sky (orbital parameters)

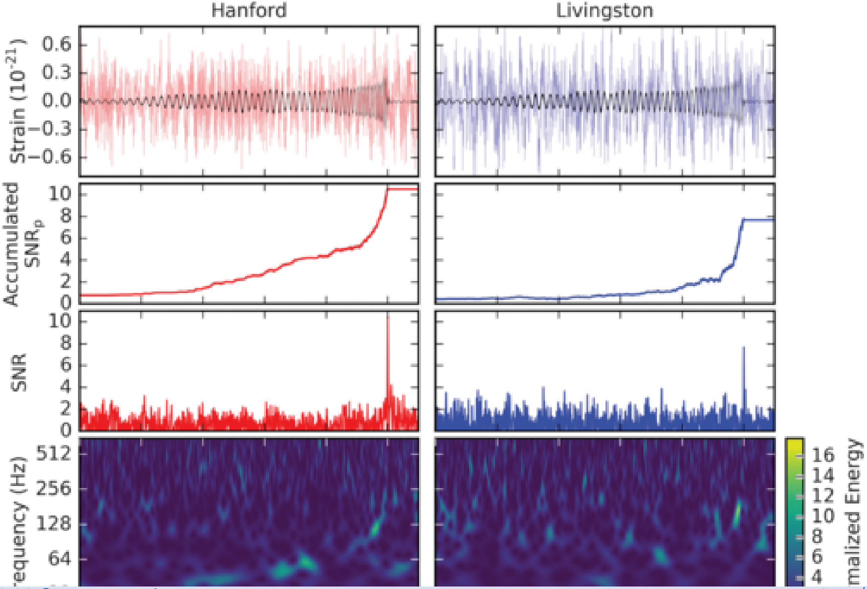
The data D

Extended example of inference: LIGO



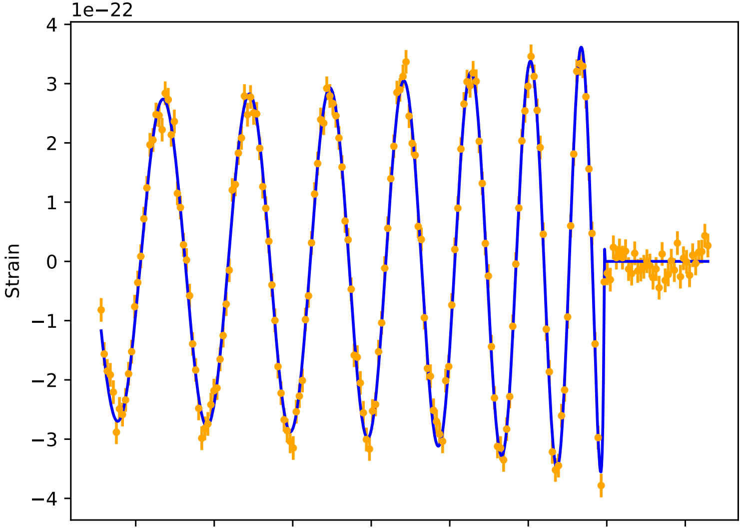
The data D

Extended example of inference: LIGO



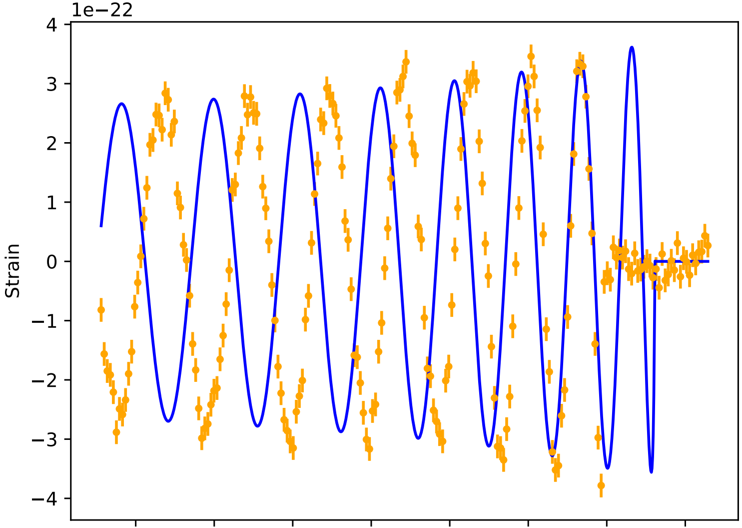
The Likelihood: well matched

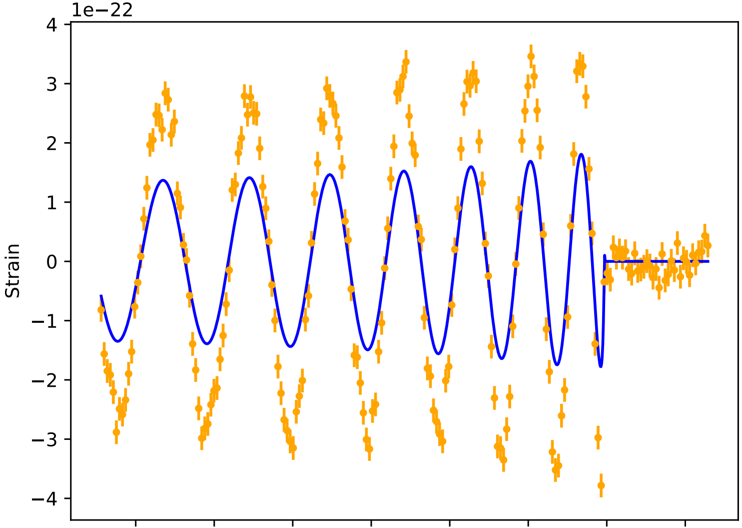
Extended example of inference: LIGO



The Likelihood: coalescence off

Extended example of inference: LIGO

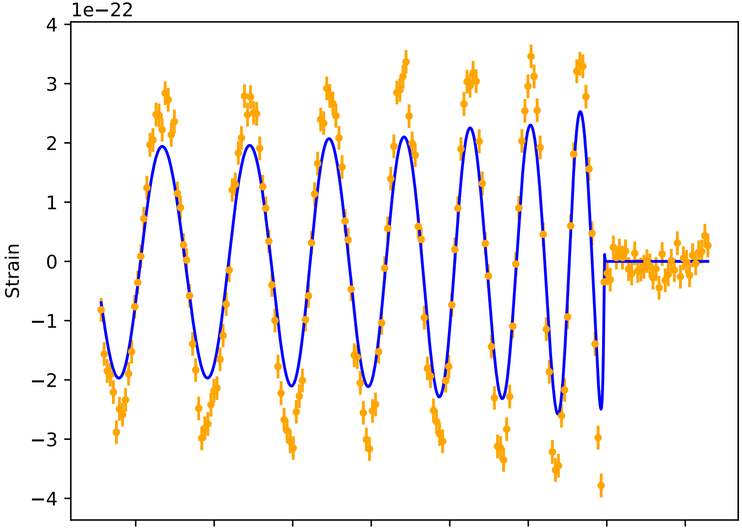




1e−22
4
3
2
1
0
−1
−2
−3
−4
Strain

The Likelihood: incorrect inclination

Extended example of inference: LIGO



- Cannot plot the full posterior distribution:

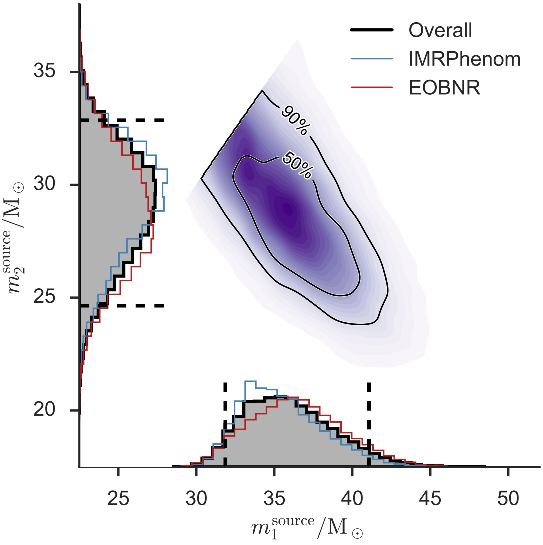
$$\mathcal{P}(\Theta) \equiv P(m_1, m_2, \theta, \phi, r, \Phi_c, t_c, i, \theta_{\mathrm{sky}}|D, M)$$

- Can plot 1D and 2D *marginalised* distributions e.g:

$$P(m_1, m_2|D, M) =$$
$$\int P(m_1, m_2, \theta, \phi, r, \Phi_c, t_c, i, \theta_{\mathrm{sky}}|D, M)\, d\theta\, d\phi\, dr\, d\Phi_c\, dt_c\, di\, d\theta_{\mathrm{sky}}$$

Posterior $\mathcal{P}$

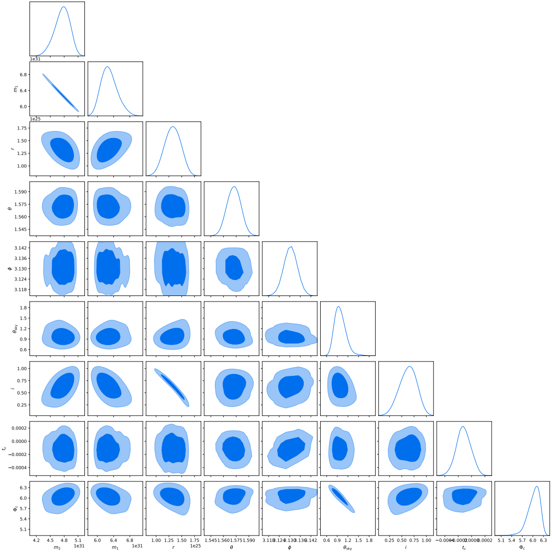
Extended example of inference: LIGO



- May do this for each pair of parameters
- Generates a *triangle plot*

Posterior $\mathcal{P}$

Extended example of inference: LIGO



- Does give insight
- Not the full picture

Evidences and model comparison

Extended example of inference: LIGO

- Up until now, we have discussed *Parameter estimation*: inferring what data tell us about parameters Θ of a model M.
- Scientifically speaking, this is only half the story.
- In general, we will have several competing models that describe the data, and we want to know which is the "best".

Parameter estimation

Another example.

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

Parameter estimation

Another example.

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

$$D = \{C_\ell\}$$

Parameter estimation

Another example.

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

$$\begin{aligned} D &= \{C_\ell\} \\ M &= \Lambda\mathrm{CDM} \end{aligned}$$

Parameter estimation

Another example.

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

$$\begin{aligned} D &= \{C_\ell\} \\ M &= \Lambda\mathrm{CDM} \\ \Theta &= \Theta_{\Lambda\mathrm{CDM}} \end{aligned}$$

Parameter estimation

Another example.

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

$$\begin{aligned} D =& \{C_\ell\} \\ M =& \Lambda\text{CDM} \\ \Theta =& \Theta_{\Lambda\text{CDM}} \\ \Theta_{\Lambda\text{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10} A_s), n_s) \end{aligned}$$

Parameter estimation

Another example.

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

$$\begin{aligned}
D =& \{C_\ell^{(\text{Planck})}\} \\
M =& \Lambda\text{CDM} \\
\Theta =& \Theta_{\Lambda\text{CDM}} \\
\Theta_{\Lambda\text{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10} A_s), n_s)
\end{aligned}$$

Parameter estimation

Another example.

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

$$\begin{aligned}
D =& \{C_\ell^{(\mathrm{Planck})}\} \\
M =& \Lambda\mathrm{CDM} \\
\Theta =& \Theta_{\Lambda\mathrm{CDM}} + \Theta_{\mathrm{Planck}} \\
\Theta_{\Lambda\mathrm{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10} A_s), n_s)
\end{aligned}$$

Parameter estimation

Another example.

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

$$\begin{aligned}
D =& \{C_\ell^{(\text{Planck})}\} \\
M =& \Lambda\text{CDM} \\
\Theta =& \Theta_{\Lambda\text{CDM}} + \Theta_{\text{Planck}} \\
\Theta_{\Lambda\text{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10} A_s), n_s) \\
\Theta_{\text{Planck}} =& (y_{\mathrm{cal}}, A^{CIB}_{217}, \xi^{tSZ-CIB}, A^{tSZ}_{143}, A^{PS}_{100}, A^{PS}_{143}, A^{PS}_{143\times217}, A^{PS}_{217}, \\
& A^{kSZ}, A^{\mathrm{dust}\,TT}_{100}, A^{\mathrm{dust}\,TT}_{143}, A^{\mathrm{dust}\,TT}_{143\times217}, A^{\mathrm{dust}\,TT}_{217}, c_{100}, c_{217})
\end{aligned}$$

Parameter estimation

Another example.

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

$$\begin{aligned}
D =& \{C_\ell^{\text{(Planck)}}\} \\
M =& \Lambda\text{CDM} + \text{extensions} \\
\Theta =& \Theta_{\Lambda\text{CDM}} + \Theta_{\text{Planck}} \\
\Theta_{\Lambda\text{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10} A_s), n_s) \\
\Theta_{\text{Planck}} =& (y_{\mathrm{cal}}, A_{217}^{CIB}, \xi^{tSZ-CIB}, A_{143}^{tSZ}, A_{100}^{PS}, A_{143}^{PS}, A_{143\times217}^{PS}, A_{217}^{PS}, \\
& A^{kSZ}, A_{100}^{\mathrm{dust}\,TT}, A_{143}^{\mathrm{dust}\,TT}, A_{143\times217}^{\mathrm{dust}\,TT}, A_{217}^{\mathrm{dust}\,TT}, c_{100}, c_{217})
\end{aligned}$$

Parameter estimation

Another example.

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

$$\begin{aligned}
D =& \{ C_\ell^{(\text{Planck})} \} \\
M =& \Lambda\text{CDM} + \text{extensions} \\
\Theta =& \Theta_{\Lambda\text{CDM}} + \Theta_{\text{Planck}} + \Theta_{\text{extensions}} \\
\Theta_{\Lambda\text{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10} A_s), n_s) \\
\Theta_{\text{Planck}} =& (y_{\mathrm{cal}}, A^{CIB}_{217}, \xi^{tSZ-CIB}, A^{tSZ}_{143}, A^{PS}_{100}, A^{PS}_{143}, A^{PS}_{143\times217}, A^{PS}_{217}, \\
& A^{kSZ}, A^{\mathrm{dust}\,TT}_{100}, A^{\mathrm{dust}\,TT}_{143}, A^{\mathrm{dust}\,TT}_{143\times217}, A^{\mathrm{dust}\,TT}_{217}, c_{100}, c_{217})
\end{aligned}$$

Parameter estimation

Another example.

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

$$\begin{aligned}
D =& \{ C_\ell^{(\text{Planck})} \} \\
M =& \Lambda\text{CDM} + \text{extensions} \\
\Theta =& \Theta_{\Lambda\text{CDM}} + \Theta_{\text{Planck}} + \Theta_{\text{extensions}} \\
\Theta_{\Lambda\text{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10} A_s), n_s) \\
\Theta_{\text{Planck}} =& (y_{\mathrm{cal}}, A_{217}^{CIB}, \xi^{tSZ-CIB}, A_{143}^{tSZ}, A_{100}^{PS}, A_{143}^{PS}, A_{143\times 217}^{PS}, A_{217}^{PS}, \\
& A^{kSZ}, A_{100}^{\mathrm{dust}\,TT}, A_{143}^{\mathrm{dust}\,TT}, A_{143\times 217}^{\mathrm{dust}\,TT}, A_{217}^{\mathrm{dust}\,TT}, c_{100}, c_{217}) \\
\Theta_{\text{extensions}} =& (n_{\mathrm{run}})
\end{aligned}$$

Parameter estimation

Another example.

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

$$\begin{aligned}
D =& \{C_\ell^{(\text{Planck})}\} \\
M =& \Lambda\text{CDM} + \text{extensions} \\
\Theta =& \Theta_{\Lambda\text{CDM}} + \Theta_{\text{Planck}} + \Theta_{\text{extensions}} \\
\Theta_{\Lambda\text{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10} A_s), n_s) \\
\Theta_{\text{Planck}} =& (y_{\rm cal}, A^{CIB}_{217}, \xi^{tSZ-CIB}, A^{tSZ}_{143}, A^{PS}_{100}, A^{PS}_{143}, A^{PS}_{143\times 217}, A^{PS}_{217}, \\
& A^{kSZ}, A^{{\rm dust}\,TT}_{100}, A^{{\rm dust}\,TT}_{143}, A^{{\rm dust}\,TT}_{143\times 217}, A^{{\rm dust}\,TT}_{217}, c_{100}, c_{217}) \\
\Theta_{\text{extensions}} =& (n_{\rm run}, n_{\rm run,run})
\end{aligned}$$

Parameter estimation

Another example.

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

$$\begin{aligned}
D =& \{C_\ell^{(\text{Planck})}\} \\
M =& \Lambda\text{CDM} + \text{extensions} \\
\Theta =& \Theta_{\Lambda\text{CDM}} + \Theta_{\text{Planck}} + \Theta_{\text{extensions}} \\
\Theta_{\Lambda\text{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10}A_s), n_s) \\
\Theta_{\text{Planck}} =& (y_{\mathrm{cal}}, A_{217}^{CIB}, \xi^{tSZ-CIB}, A_{143}^{tSZ}, A_{100}^{PS}, A_{143}^{PS}, A_{143\times217}^{PS}, A_{217}^{PS}, \\
& A^{kSZ}, A_{100}^{\mathrm{dust}\,TT}, A_{143}^{\mathrm{dust}\,TT}, A_{143\times217}^{\mathrm{dust}\,TT}, A_{217}^{\mathrm{dust}\,TT}, c_{100}, c_{217}) \\
\Theta_{\text{extensions}} =& (n_{\mathrm{run}}, n_{\mathrm{run,run}}, w)
\end{aligned}$$

Parameter estimation

Another example.

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

$$\begin{aligned}
D =& \{ C_\ell^{(\text{Planck})} \} \\
M =& \Lambda\text{CDM} + \text{extensions} \\
\Theta =& \Theta_{\Lambda\text{CDM}} + \Theta_{\text{Planck}} + \Theta_{\text{extensions}} \\
\Theta_{\Lambda\text{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10} A_s), n_s) \\
\Theta_{\text{Planck}} =& (y_{\mathrm{cal}}, A^{CIB}_{217}, \xi^{tSZ-CIB}, A^{tSZ}_{143}, A^{PS}_{100}, A^{PS}_{143}, A^{PS}_{143\times217}, A^{PS}_{217}, \\
& A^{kSZ}, A^{\mathrm{dust}\,TT}_{100}, A^{\mathrm{dust}\,TT}_{143}, A^{\mathrm{dust}\,TT}_{143\times217}, A^{\mathrm{dust}\,TT}_{217}, c_{100}, c_{217}) \\
\Theta_{\text{extensions}} =& (n_{\mathrm{run}}, n_{\mathrm{run,run}}, w, \Sigma m_\nu, m^{\mathrm{eff}}_{\nu,\mathrm{sterile}})
\end{aligned}$$

Parameter estimation

Another example.

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

$$\begin{aligned}
D =& \{C_\ell^{(\text{Planck})}\} + \{\text{LSS}\} \\
M =& \Lambda\text{CDM} + \text{extensions} \\
\Theta =& \Theta_{\Lambda\text{CDM}} + \Theta_{\text{Planck}} + \Theta_{\text{extensions}} \\
\Theta_{\Lambda\text{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10} A_s), n_s) \\
\Theta_{\text{Planck}} =& (y_{\mathrm{cal}}, A_{217}^{CIB}, \xi^{tSZ-CIB}, A_{143}^{tSZ}, A_{100}^{PS}, A_{143}^{PS}, A_{143\times217}^{PS}, A_{217}^{PS}, \\
& A^{kSZ}, A_{100}^{\mathrm{dust}\,TT}, A_{143}^{\mathrm{dust}\,TT}, A_{143\times217}^{\mathrm{dust}\,TT}, A_{217}^{\mathrm{dust}\,TT}, c_{100}, c_{217}) \\
\Theta_{\text{extensions}} =& (n_{\mathrm{run}}, n_{\mathrm{run,run}}, w, \Sigma m_\nu, m_{\nu,\mathrm{sterile}}^{\mathrm{eff}})
\end{aligned}$$

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

$$\begin{aligned}
D =& \{C_\ell^{(\text{Planck})}\} + \{\text{LSS}\} + \{\text{"Big Data"}\} \\
M =& \Lambda\text{CDM} + \text{extensions} \\
\Theta =& \Theta_{\Lambda\text{CDM}} + \Theta_{\text{Planck}} + \Theta_{\text{extensions}} \\
\Theta_{\Lambda\text{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10} A_s), n_s) \\
\Theta_{\text{Planck}} =& (y_{\mathrm{cal}}, A_{217}^{CIB}, \xi^{tSZ-CIB}, A_{143}^{tSZ}, A_{100}^{PS}, A_{143}^{PS}, A_{143\times 217}^{PS}, A_{217}^{PS}, \\
& A^{kSZ}, A_{100}^{\mathrm{dust}\,TT}, A_{143}^{\mathrm{dust}\,TT}, A_{143\times 217}^{\mathrm{dust}\,TT}, A_{217}^{\mathrm{dust}\,TT}, c_{100}, c_{217}) \\
\Theta_{\text{extensions}} =& (n_{\mathrm{run}}, n_{\mathrm{run,run}}, w, \Sigma m_\nu, m_{\nu,\mathrm{sterile}}^{\mathrm{eff}})
\end{aligned}$$

Parameter estimation

Another example.

$$\mathcal{L}(\Theta) = P(D|\Theta, M)$$

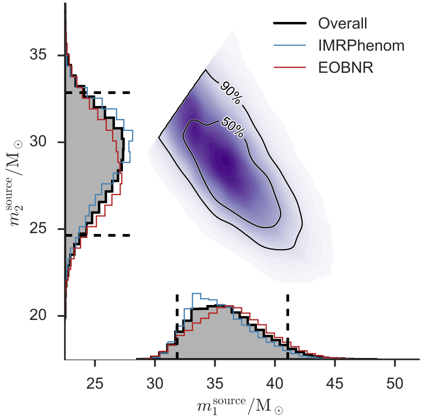
$$\begin{aligned}
D =& \{C_\ell^{(\text{Planck})}\} + \{\text{LSS}\} + \{\text{“Big Data”}\} \\
M =& \Lambda\text{CDM} + \text{extensions} \\
\Theta =& \Theta_{\Lambda\text{CDM}} + \Theta_{\text{Planck}} + \Theta_{\text{extensions}} \\
\Theta_{\Lambda\text{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10}A_s), n_s) \\
\Theta_{\text{Planck}} =& (y_{\mathrm{cal}}, A^{CIB}_{217}, \xi^{tSZ-CIB}, A^{tSZ}_{143}, A^{PS}_{100}, A^{PS}_{143}, A^{PS}_{143\times 217}, A^{PS}_{217}, \\
& A^{kSZ}, A^{\mathrm{dust}\,TT}_{100}, A^{\mathrm{dust}\,TT}_{143}, A^{\mathrm{dust}\,TT}_{143\times 217}, A^{\mathrm{dust}\,TT}_{217}, c_{100}, c_{217}) \\
\Theta_{\text{extensions}} =& (n_{\mathrm{run}}, n_{\mathrm{run,run}}, w, \Sigma m_\nu, m^{\mathrm{eff}}_{\nu,\mathrm{sterile}})
\end{aligned}$$

- Parameter estimation: $L, \pi \to \mathcal{P}$: model parameters
- Model comparison: $L, \pi \to Z$: how good model is

Sampling

How to describe a high-dimensional posterior

- In high dimensions, posterior $\mathcal{P}$ occupies a vanishingly small region of the prior π.
- *Sampling* the posterior is an excellent compression scheme.



Why do sampling?

Marginalisation over the posterior

- Set of N samples $S = \{\Theta^{(i)} : i = 1, \ldots N : \; \Theta^{(i)} \sim \mathcal{P}\}$
- Mean mass:

$$\bar{m}_1 \equiv \langle m_1 \rangle_{\mathcal{P}} \approx \frac{1}{N} \sum_{i=1}^{N} m_1^{(i)}$$

- Mass covariance:

$$\mathrm{Cov}(m_1, m_2) \approx \frac{1}{N} \sum_{i=1}^{N} (m_1^{(i)} - \bar{m}_1)(m_2^{(i)} - \bar{m}_2)$$

- Marginalised samples: Just ignore the other coordinates.
- N.B. Typically have *weighted* samples

Why do sampling?

Marginalisation over the posterior

- Set of N samples $S = \{\Theta^{(i)} : i = 1, \ldots N : \ \Theta^{(i)} \sim \mathcal{P}\}$
- Mean mass:

$$\bar{m}_1 \equiv \langle m_1 \rangle_{\mathcal{P}} \approx \frac{\sum_{i=1}^N w^{(i)} m_1^{(i)}}{\sum_{i=1}^N w^{(i)}}$$

- Mass covariance:

$$\mathrm{Cov}(m_1, m_2) \approx \frac{\sum_{i=1}^N (m_1^{(i)} - \bar{m}_1)(m_2^{(i)} - \bar{m}_2)}{\sum_{i=1}^N w^{(i)}}$$

- Marginalised samples: Just ignore the other coordinates.
- N.B. Typically have *weighted* samples

Parameter estimation

- The name of the game is therefore drawing samples S from the posterior $\mathcal{P}$ with the minimum number of likelihood calls.
- Gridding is doomed to failure in high dimensions.
- Enter Metropolis Hastings.

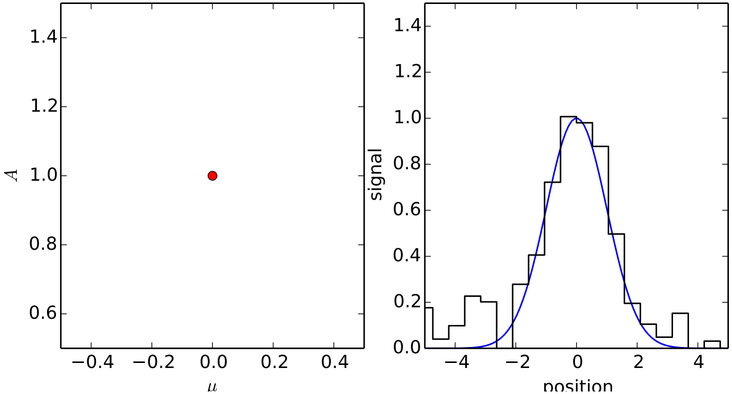
Current sampling approaches

1. Metropolis Hastings.
2. Hamiltonian Monte-Carlo (HMC).
3. Ensemble sampling (e.g. emcee).

Metropolis Hastings

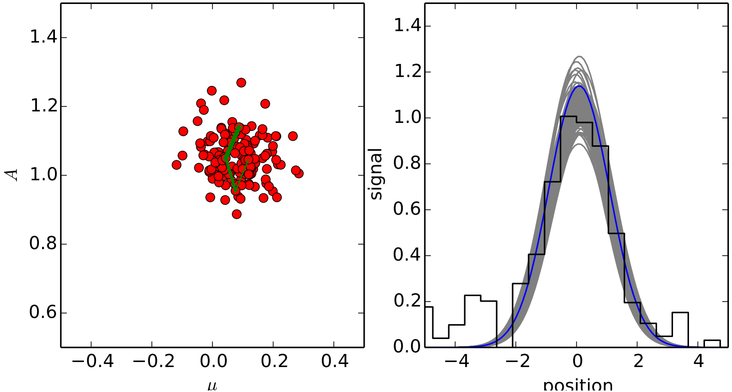
- Turn the N-dimensional problem into a one-dimensional one.
  1. Pick random direction
  2. Choose step length
  3. If uphill, make step...
  4. ...otherwise sometimes make step.

Metropolis Hastings



Metropolis Hastings

Metropolis Hastings



Metropolis Hastings

Struggles with. . .

Metropolis Hastings

Struggles with...

1. Burn in
2. Multimodality
3. Correlated Peaks
4. Phase transitions

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.
- stan is a fully fledged, rapidly developing programming language with HMC as a default sampler.

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.
- emcee is not the only (or even best) affine invariant approach.

The fundamental issue with all of the above

- They don't give you evidences!

$$\begin{aligned}\mathcal{Z} &= \mathrm{P}(D|M) \\ &= \int \mathrm{P}(D|\Theta, M)\mathrm{P}(\Theta|M) d\Theta \\ &= \langle \mathcal{L} \rangle_\pi\end{aligned}$$

- MCMC fundamentally explores the posterior, and cannot average over the prior.
- Simulated annealing gives one possibility for computing evidences.
  - Inspired by thermodynamics.
  - Suffers from similar issues to MCMC.
  - Unclear how to choose correct annealing schedule

What is nested sampling?

- Nested sampling is an alternative way of sampling posteriors.
- Uses ensemble sampling to compress prior to posterior.
- In doing so, it circumvents many issues (dimensionality, topology, geometry) that beset standard approaches.

Nested Sampling

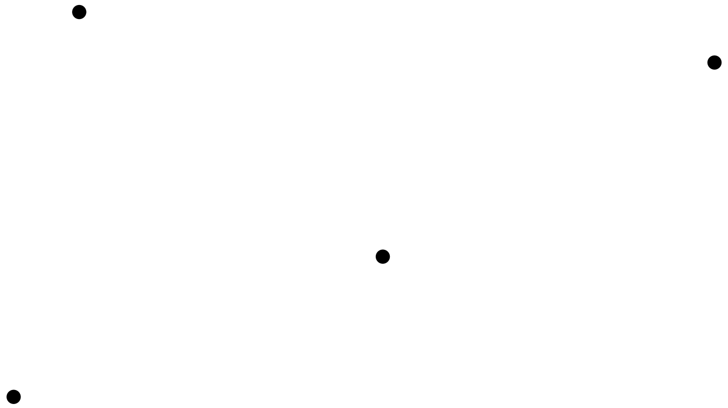
John Skilling's alternative to traditional MCMC!

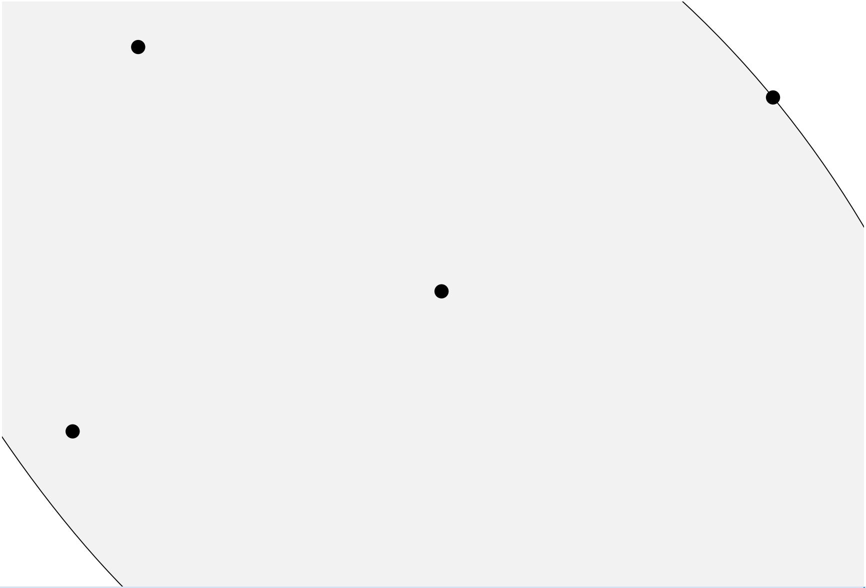
New procedure:

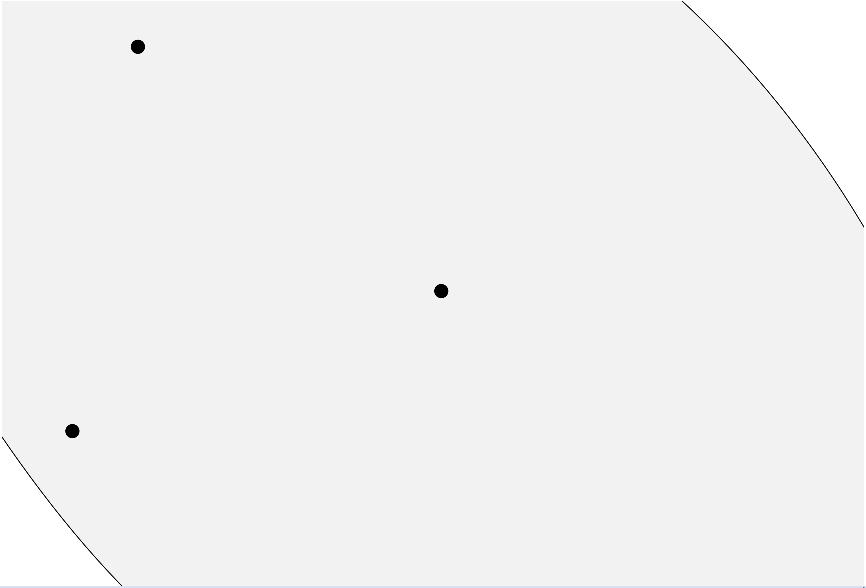
Maintain a set S of n samples, which are sequentially updated:

- S_0: Generate n samples uniformly over the space (from the prior π).
- S_{n+1}: Delete the lowest likelihood sample in S_n, and replace it with a new uniform sample with higher likelihood

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

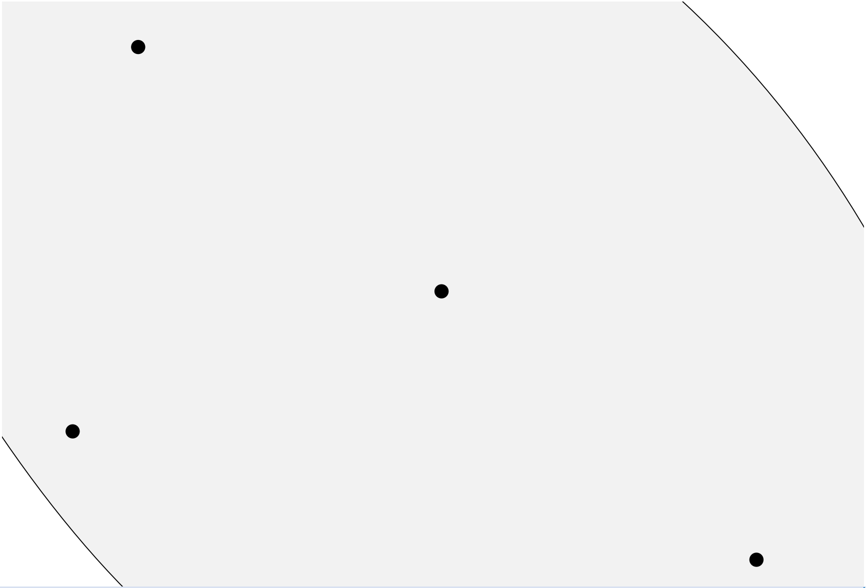


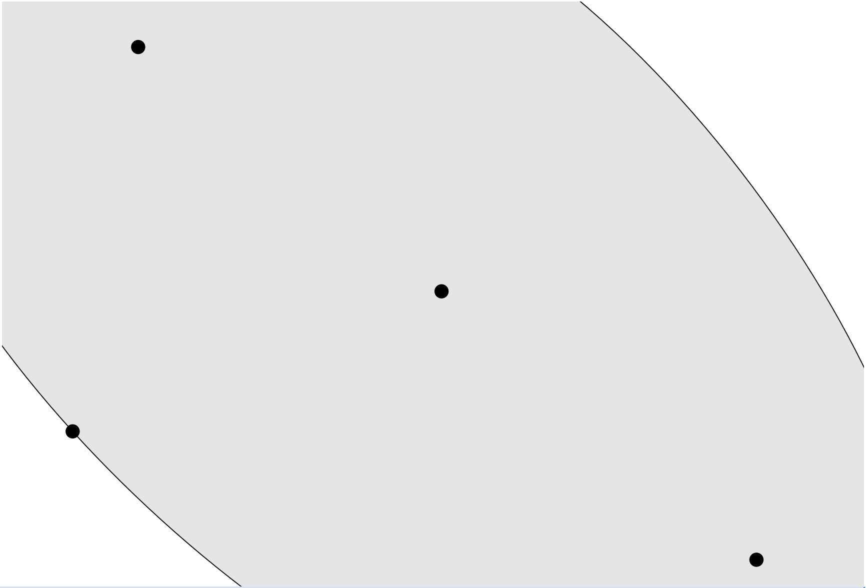




Nested Sampling

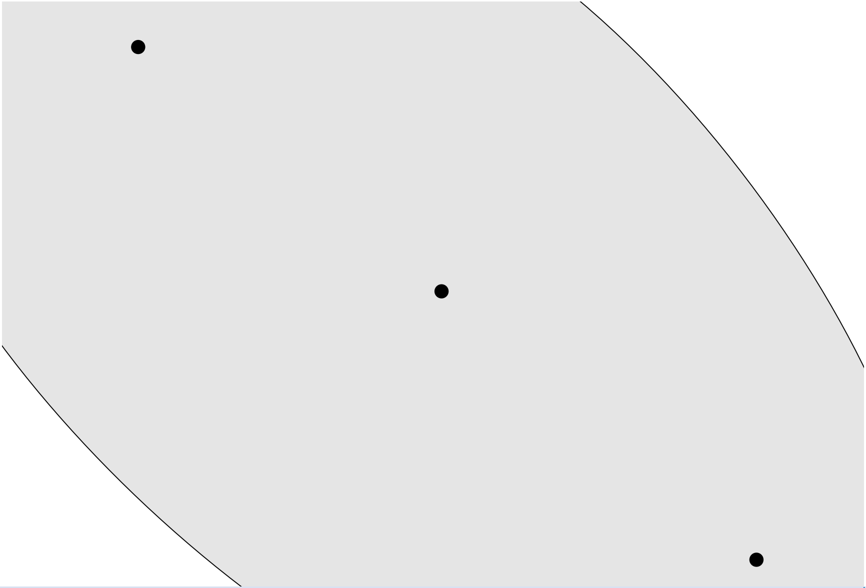
Graphical aid





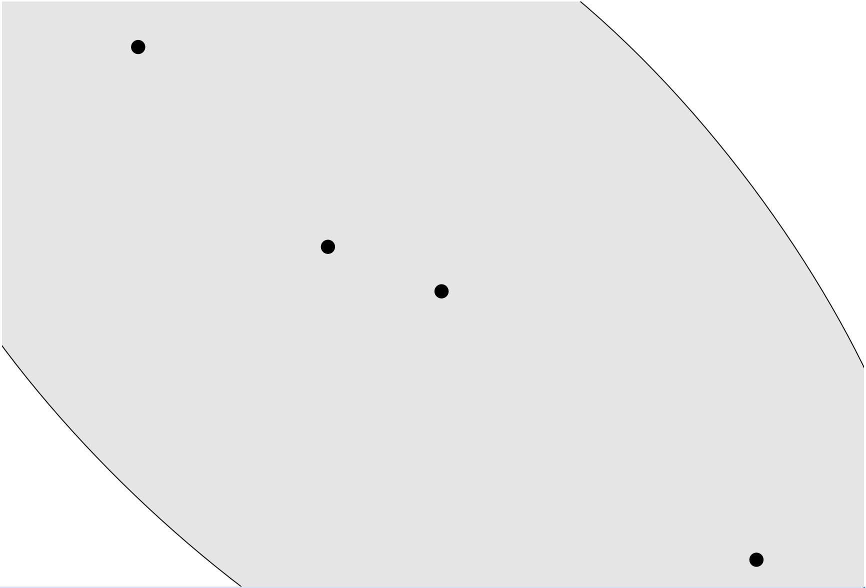
Nested Sampling

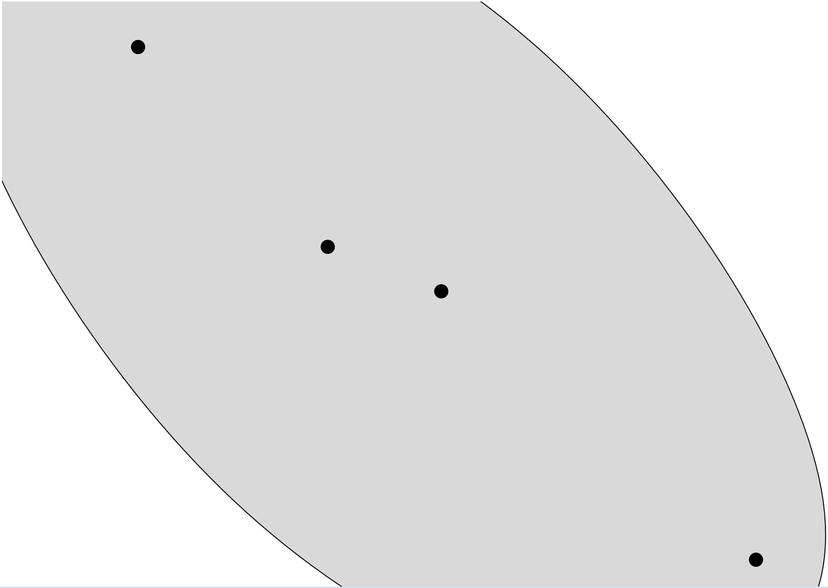
Graphical aid

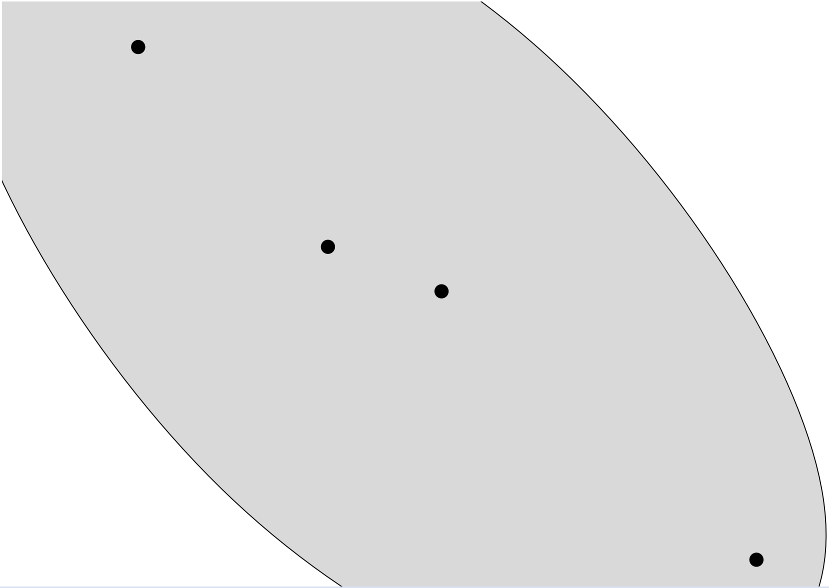


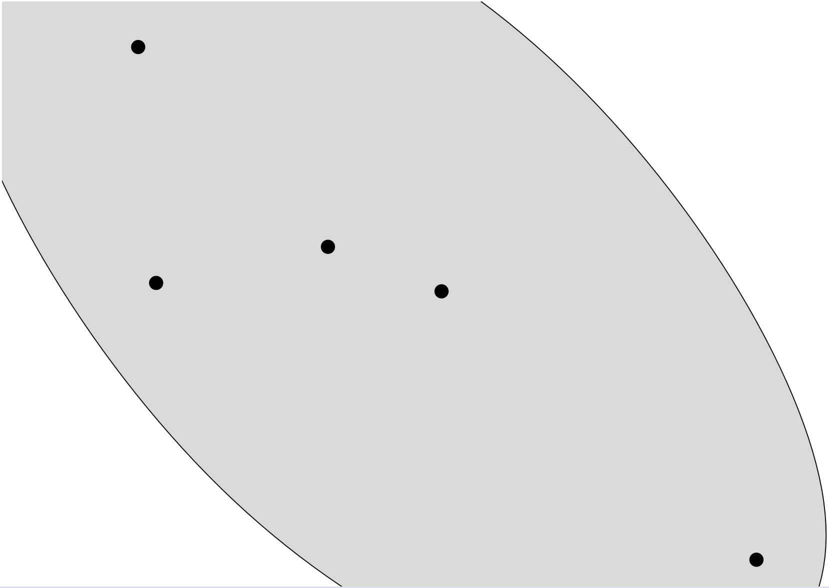
Nested Sampling

Graphical aid

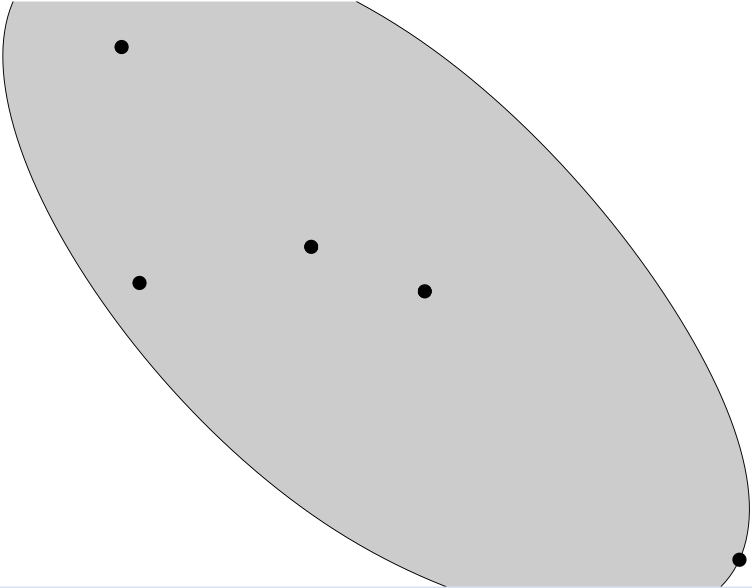






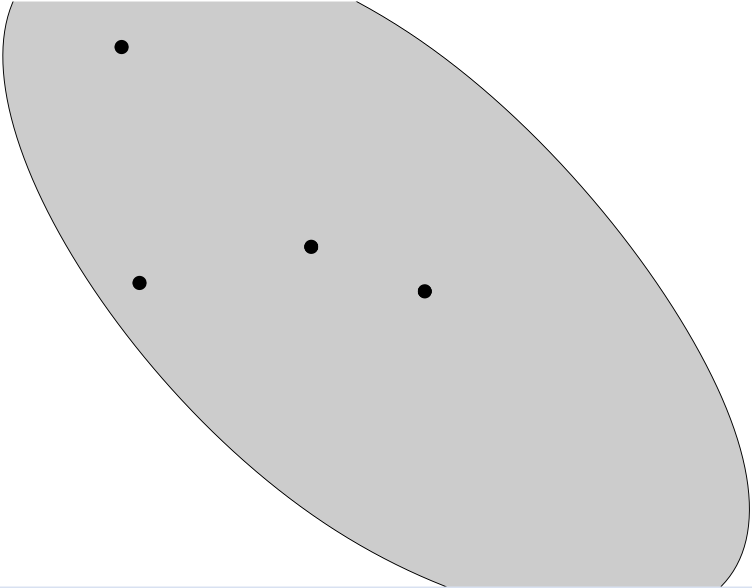


Nested Sampling
Graphical aid



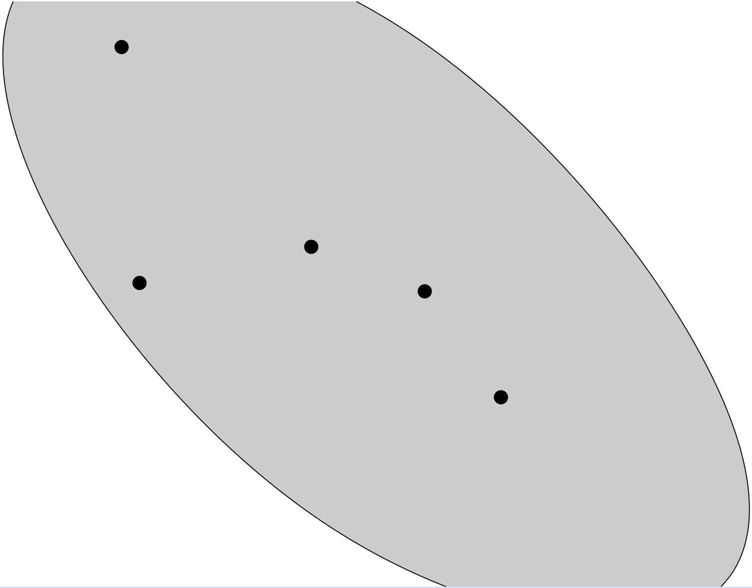
Nested Sampling

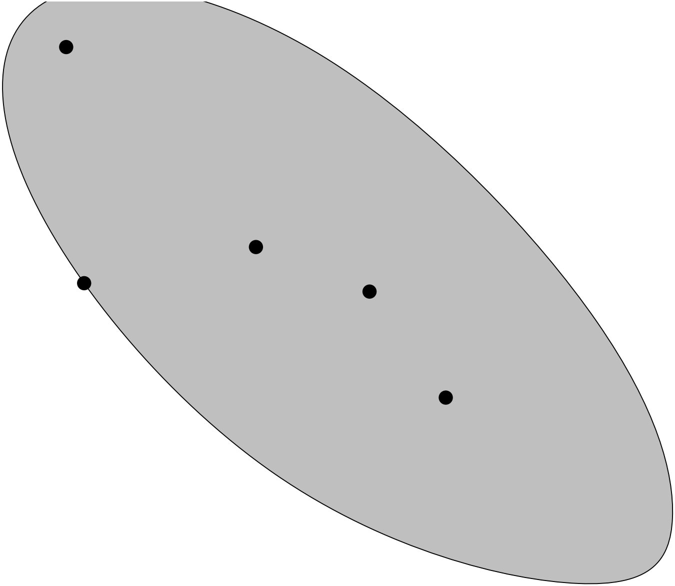
Graphical aid

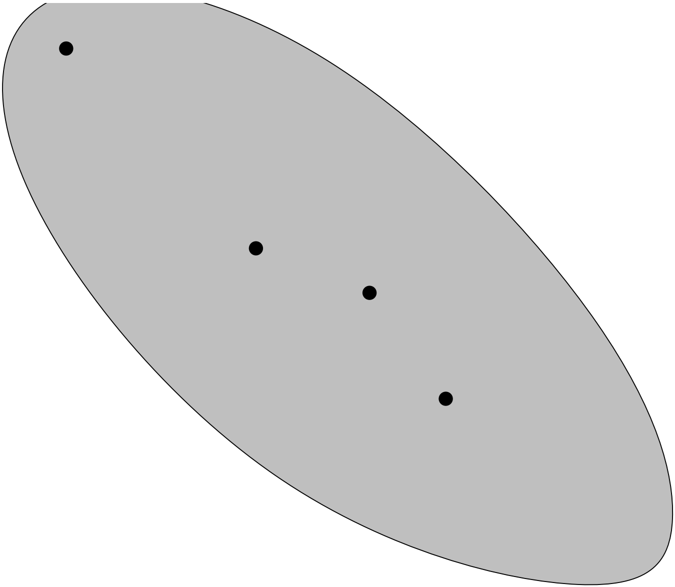


Nested Sampling

Graphical aid

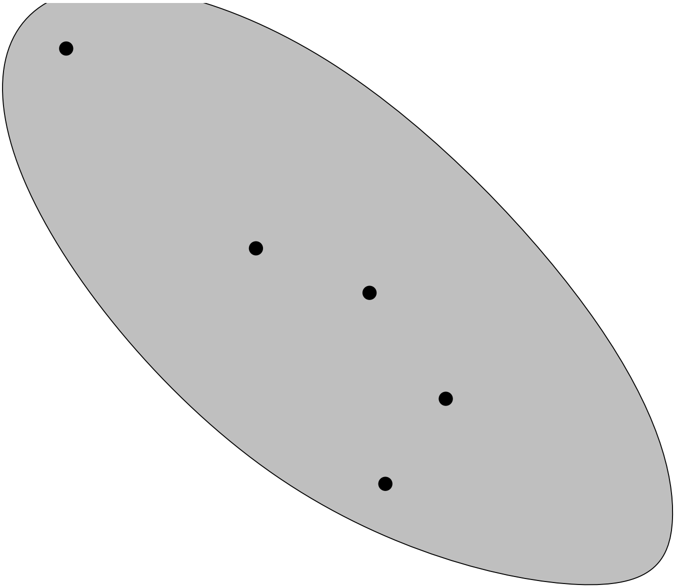


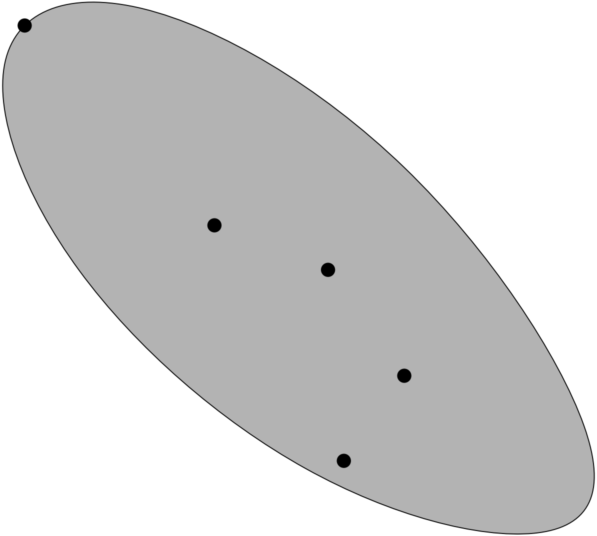




Nested Sampling

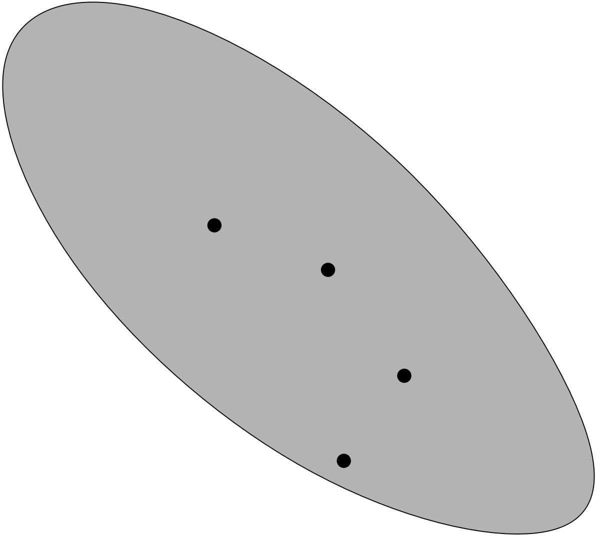
Graphical aid





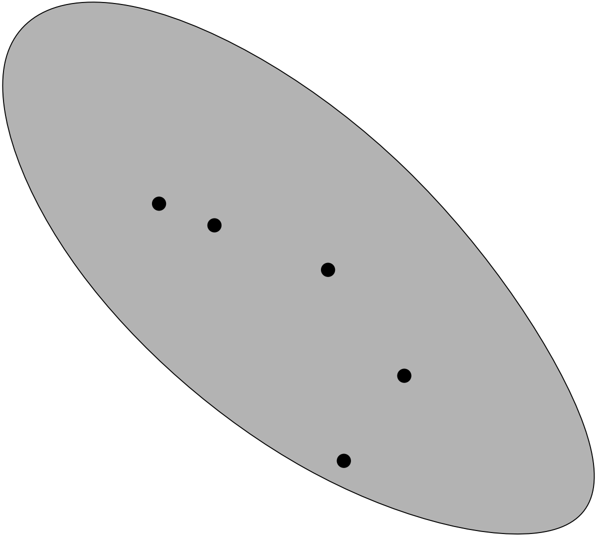
Nested Sampling

Graphical aid



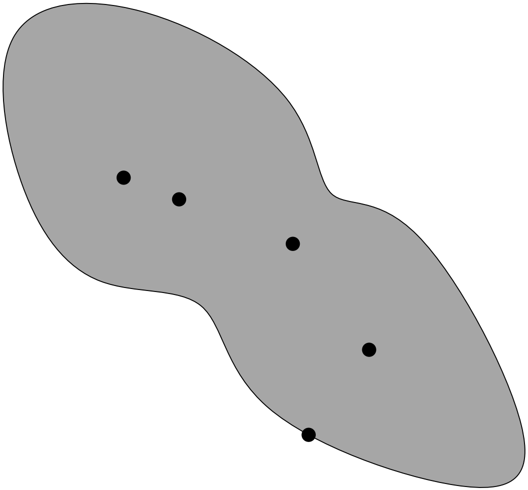
Nested Sampling

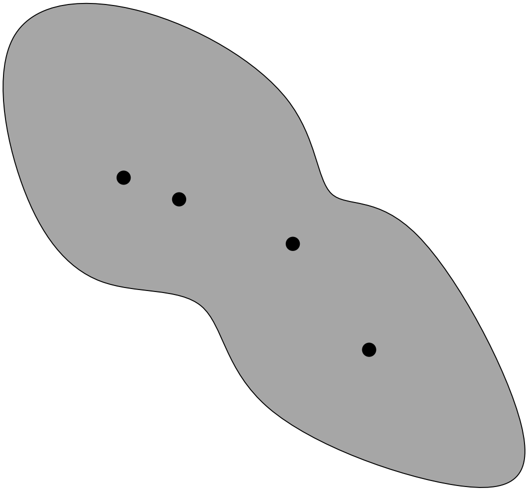
Graphical aid



Nested Sampling

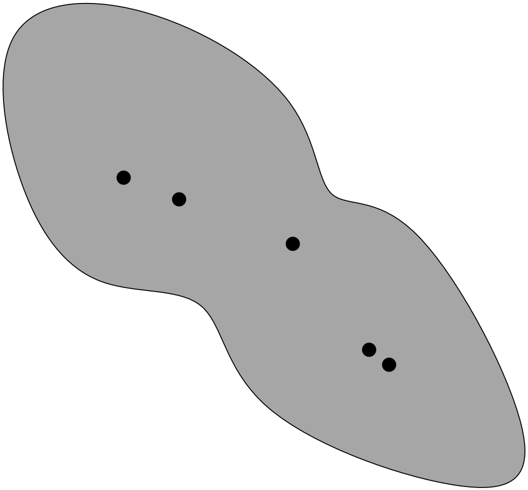
Graphical aid





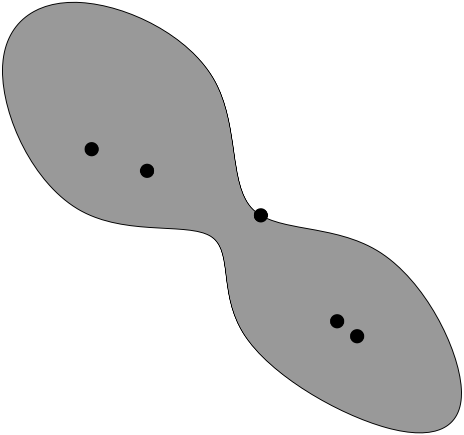
Nested Sampling

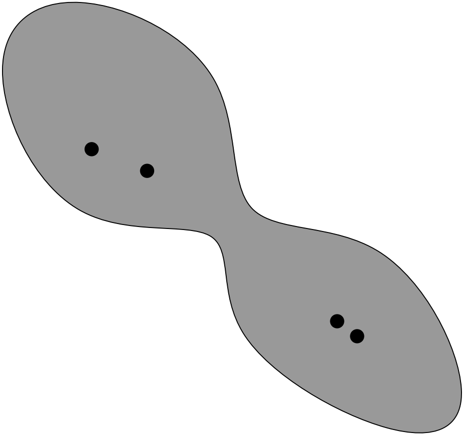
Graphical aid

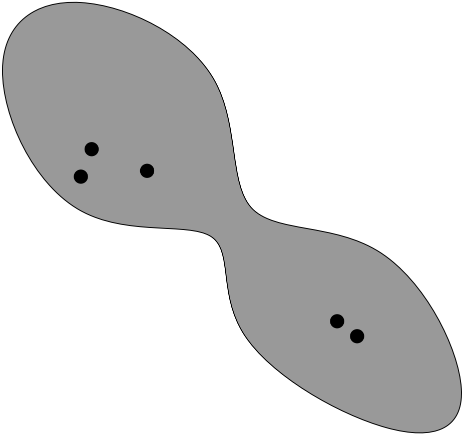


Nested Sampling

Graphical aid

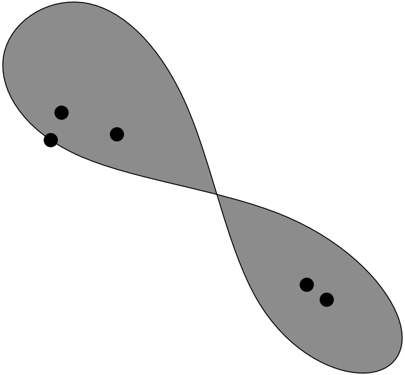


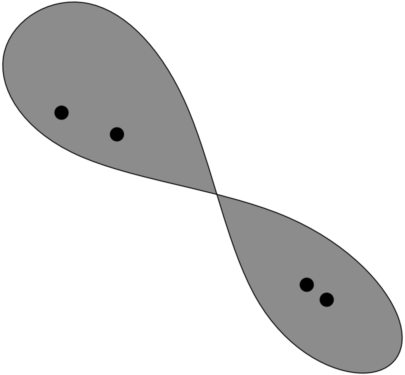




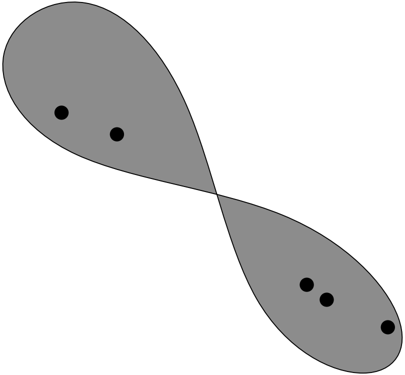
Nested Sampling

Graphical aid



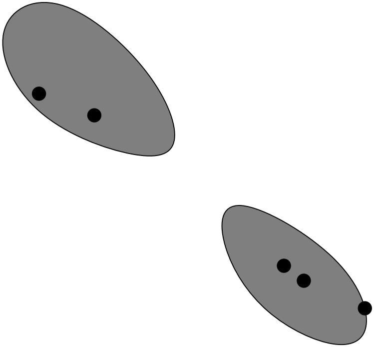


Graphical aid



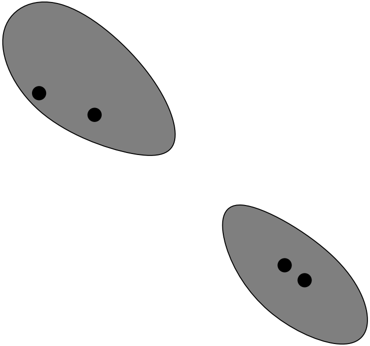
Nested Sampling

Graphical aid



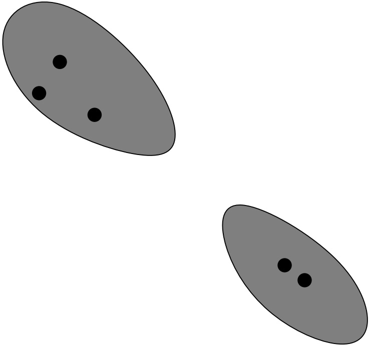
Nested Sampling

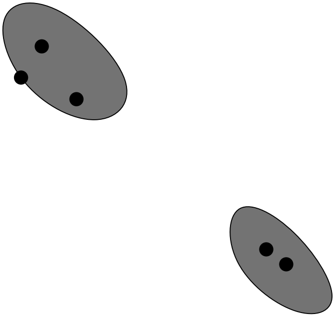
Graphical aid

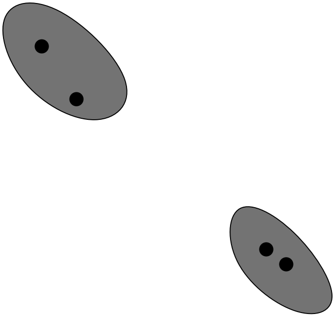


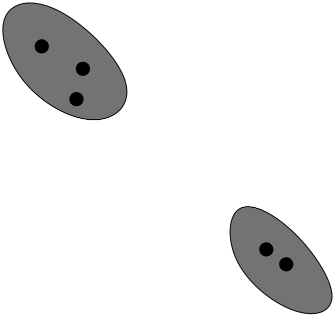
Nested Sampling

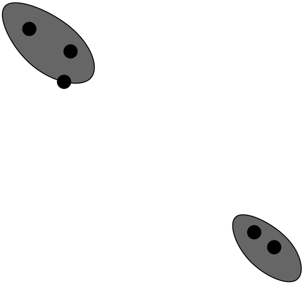
Graphical aid





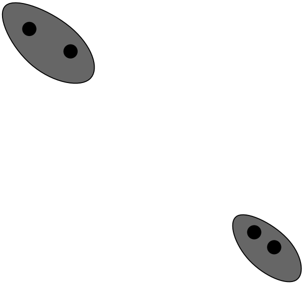


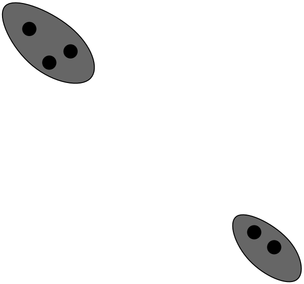




Nested Sampling

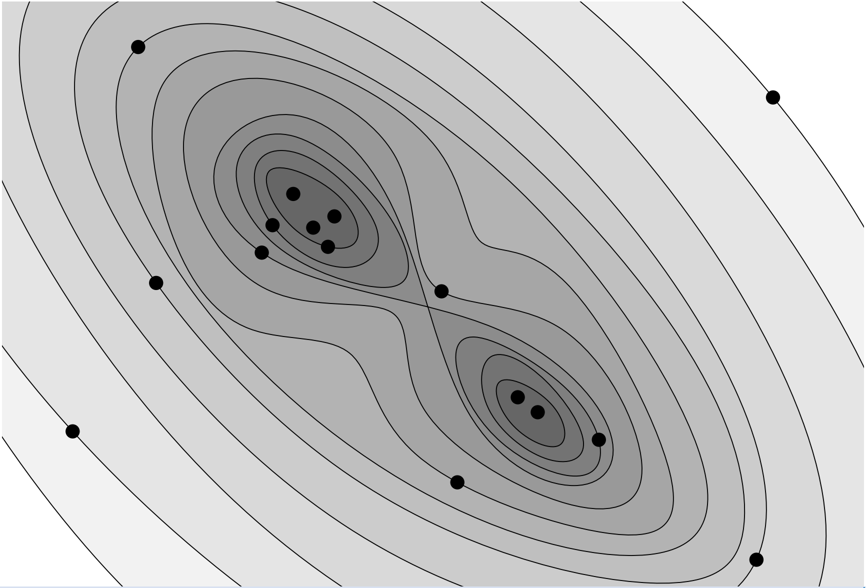
Graphical aid





Nested Sampling

Graphical aid



Nested sampling

- The set of dead points are posterior samples with an appropriate weighting factor
- They can also be used to calculate evidences, since it sequentially updates the priors.

Sampling from a hard likelihood constraint

> *"It is not the purpose of this introductory paper to develop the technology of navigation within such a volume. We merely note that exploring a hard-edged likelihood-constrained domain should prove to be neither more nor less demanding than exploring a likelihood-weighted space."*
>
> — *John Skilling*

- Most of the work in NS to date has been in attempting to implement a hard-edged sampler in the NS meta-algorithm.

Sampling within an iso-likelihood contour

Previous attempts

Rejection Sampling MultiNest; F. Feroz & M. Hobson (2009).

- Suffers in high dimensions

Hamiltonian M.J. Betancourt (2010)

Galilean F. Feroz & J. Skilling (2013)

- Requires gradients and tuning

Diffusive Nested Sampling B. Brewer et al. (2009,2016).

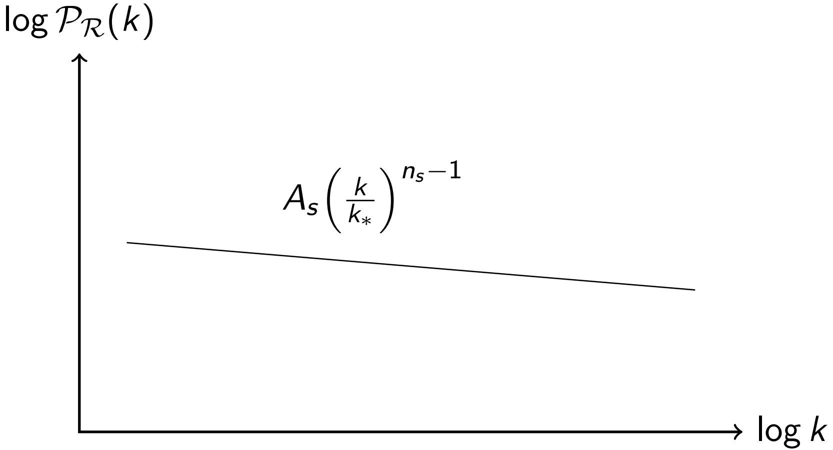
- Very promising
- Still needs tuning.

Slice Sampling PolyChord; Handley et al. (2015).

- Current "state-of-the-art".
- PolyChord 2.0 imminent.

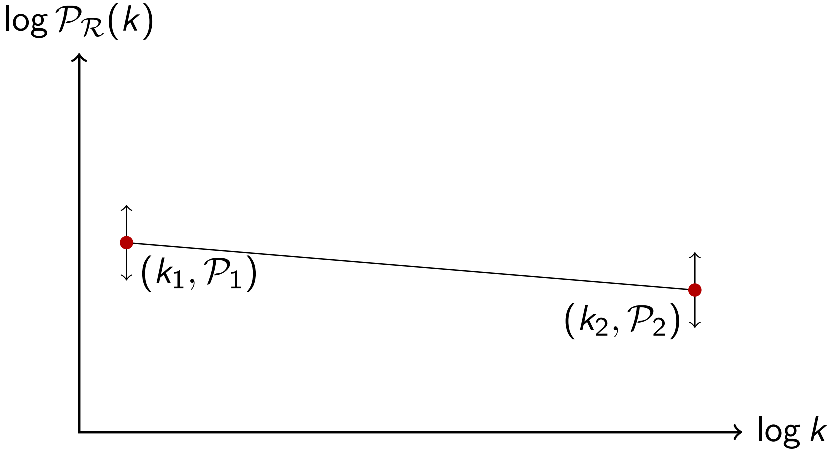
PolyChord in action

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



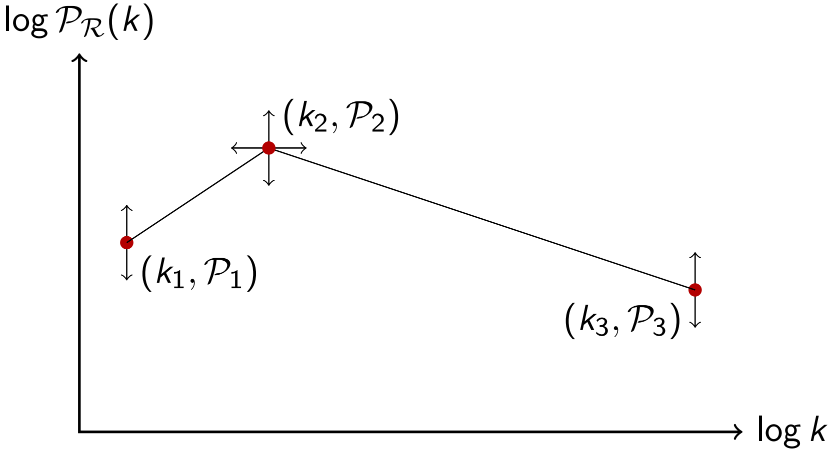
PolyChord in action

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



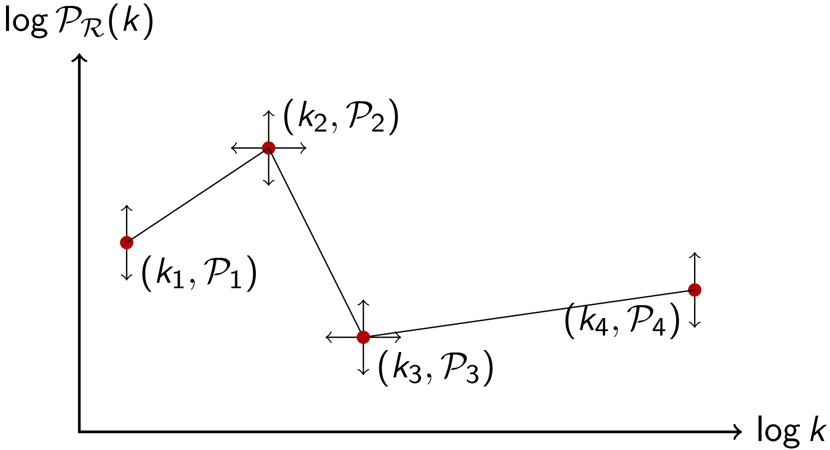
PolyChord in action

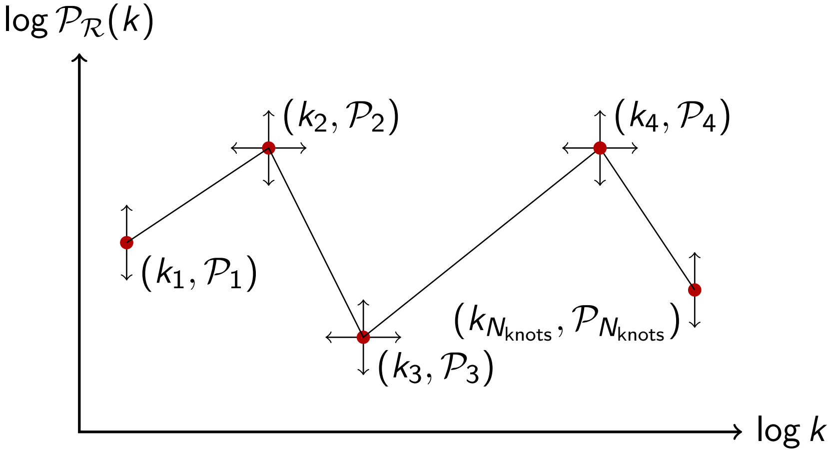
Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction



PolyChord in action

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

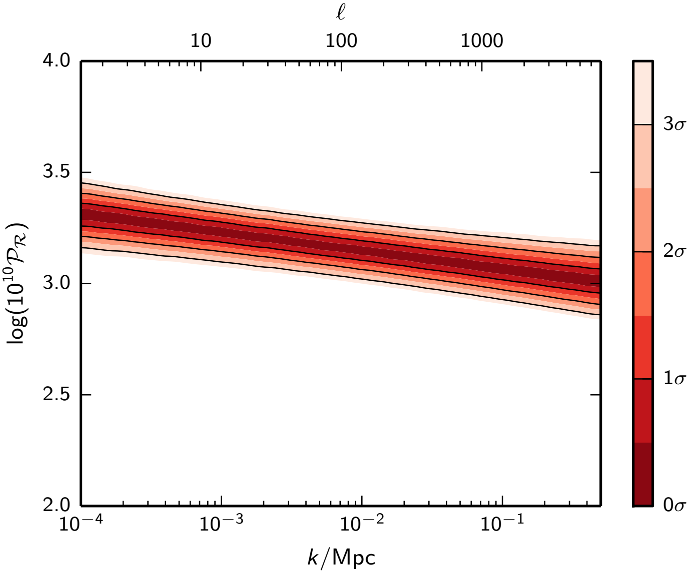




$\log \mathcal{P}_\mathcal{R}(k)$
$(k_2, \mathcal{P}_2)$
$(k_4, \mathcal{P}_4)$
$(k_1, \mathcal{P}_1)$
$(k_{N_\mathrm{knots}}, \mathcal{P}_{N_\mathrm{knots}})$
$(k_3, \mathcal{P}_3)$
$\log k$

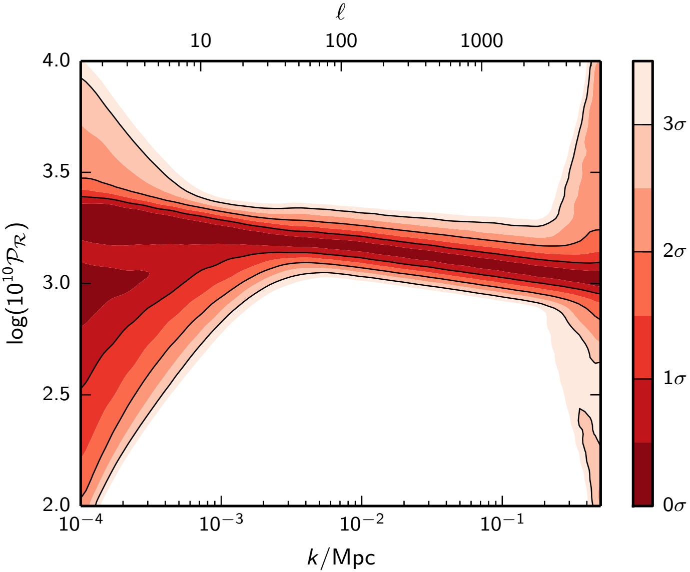
0 internal knots

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



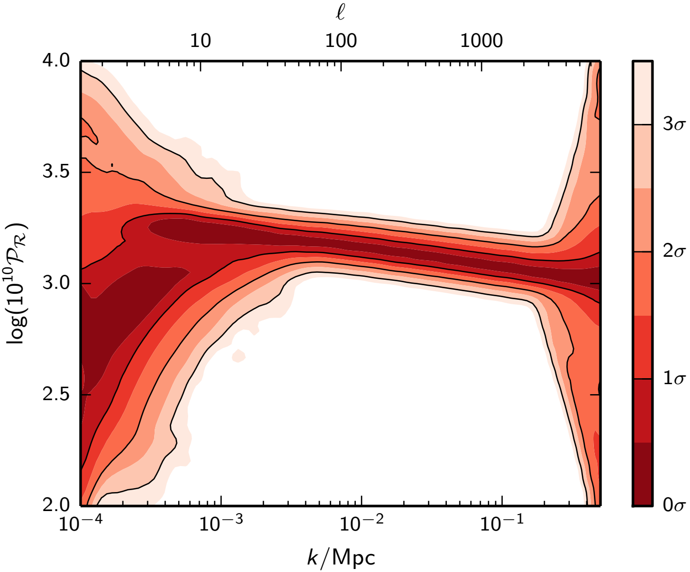
1 internal knots

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



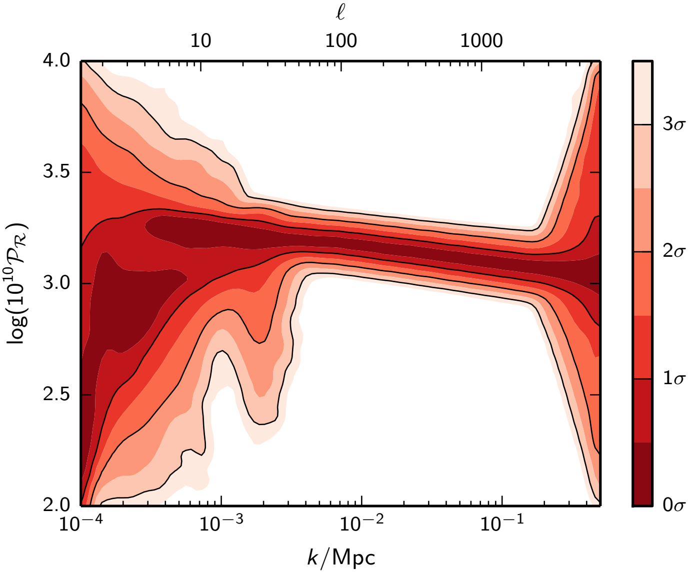
2 internal knots

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



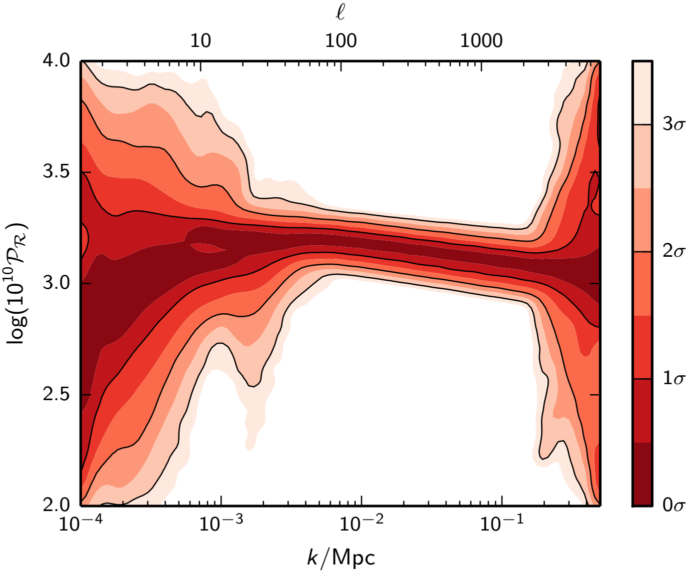
3 internal knots

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



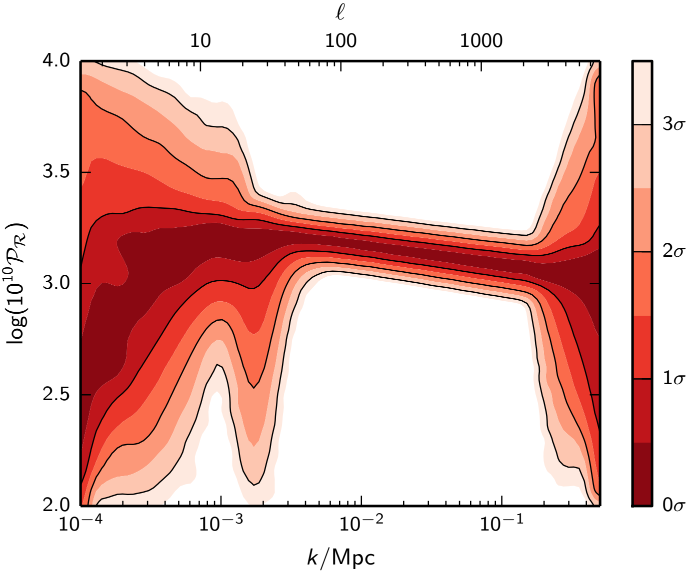
4 internal knots

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



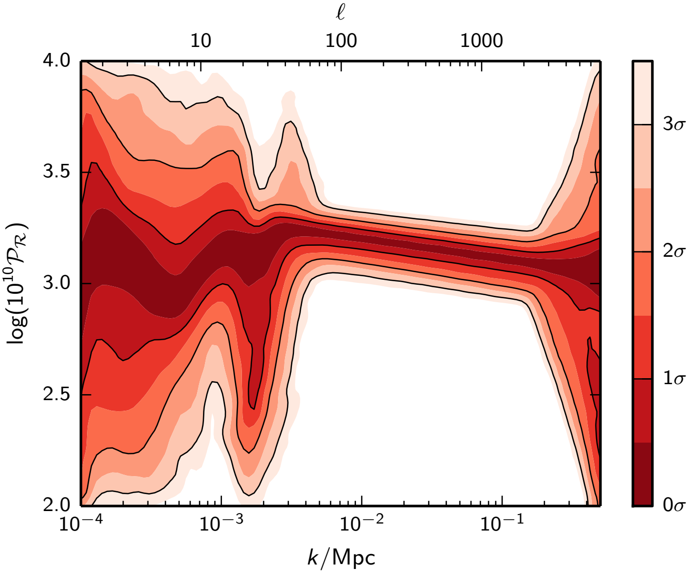
5 internal knots

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



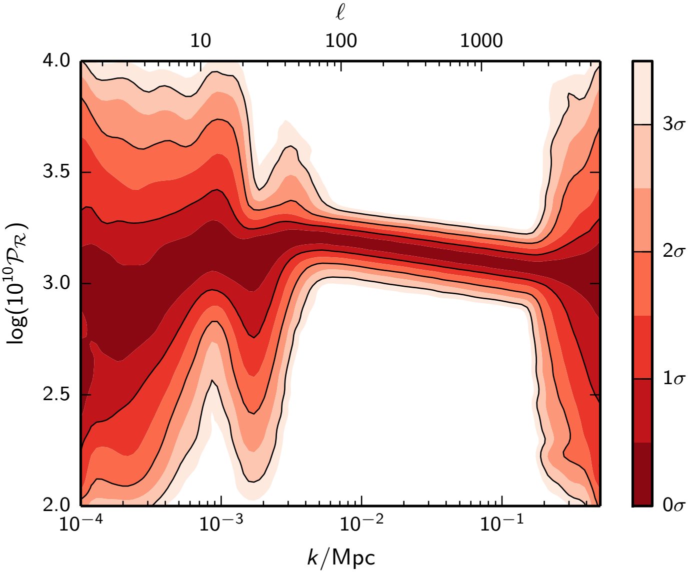
6 internal knots

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



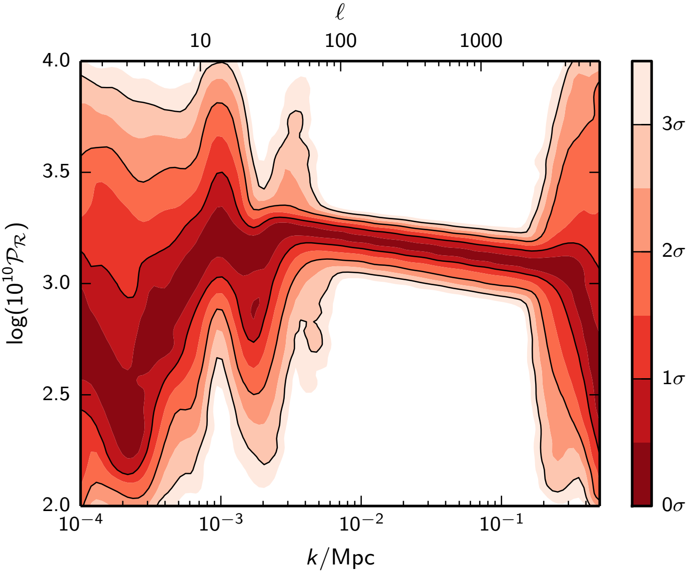
7 internal knots

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



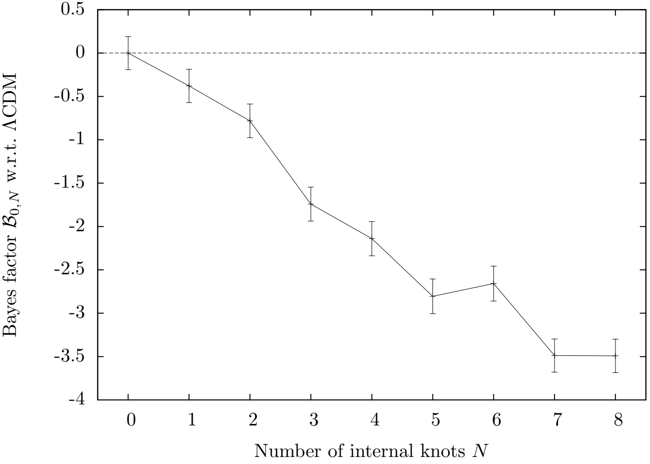
8 internal knots

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



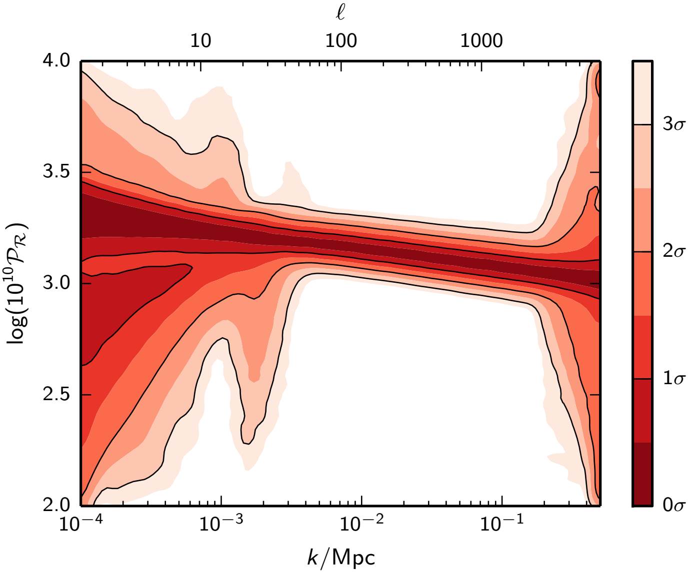
Bayes Factors

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



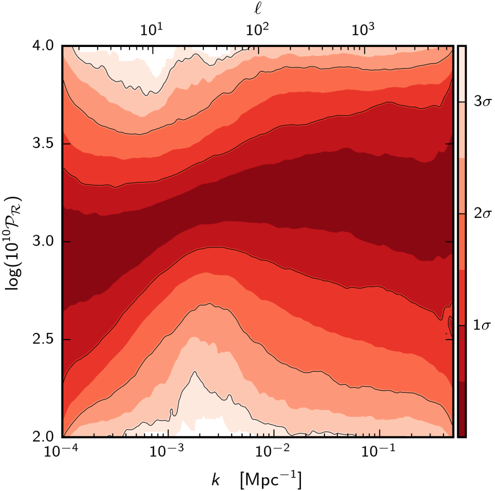
Marginalised plot

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



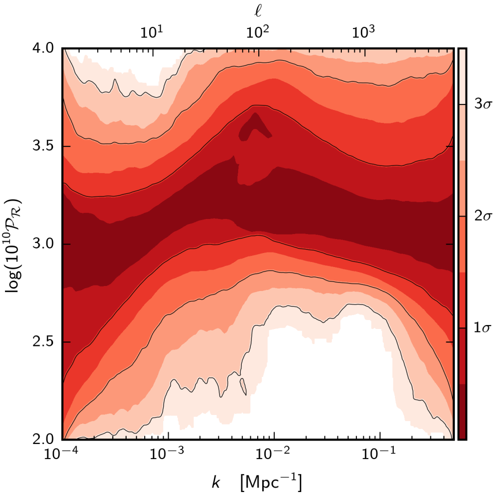
COBE (pre-2002)

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



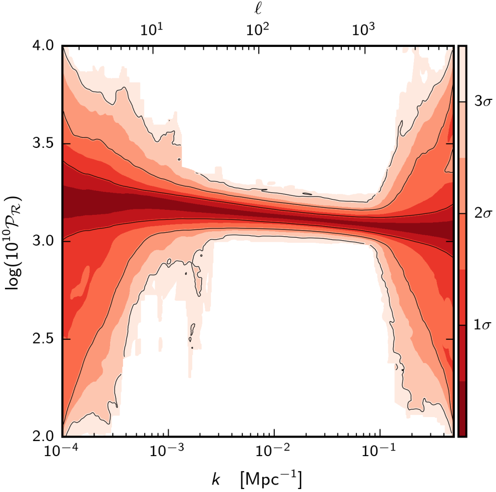
COBE et al (2002)

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



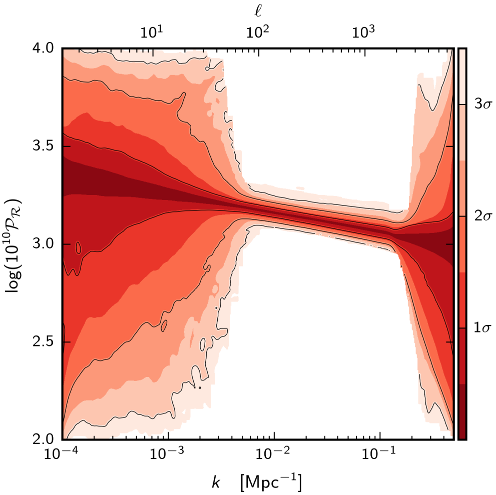
WMAP (2012)

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



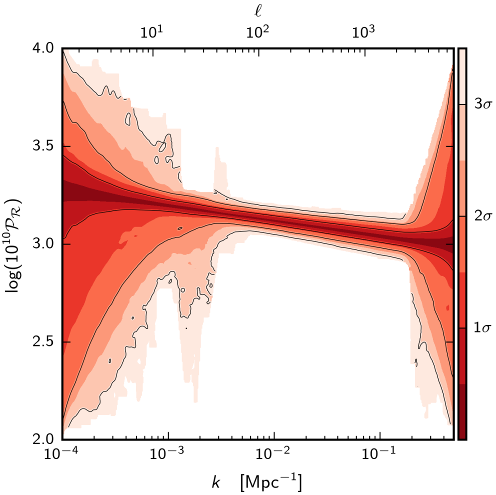
Planck (2013)

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

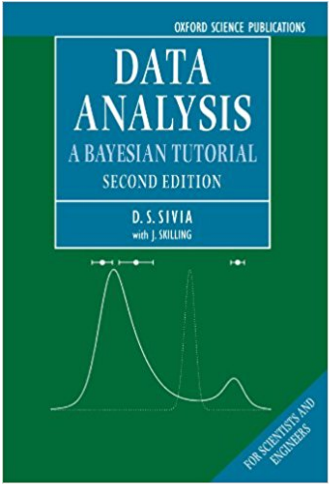


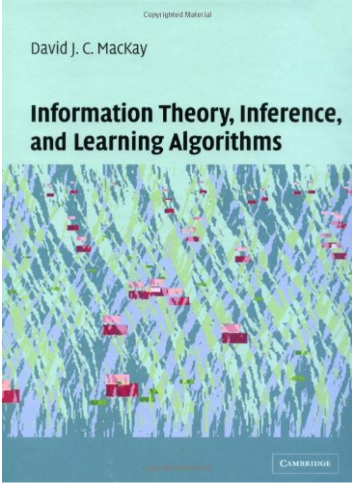
Planck (2015)

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



Further reading





- Data analysis: A Bayesian Tutorial (Sivia & Skilling)
- Information Theory, Inference and Learning Algorithms (Mackay)