Modern Bayesian Inference

Theory and Practice

Will Handley
wh260@cam.ac.uk

Astrophysics Group
Cavendish Laboratory
University of Cambridge

July 4, 2017

Introduction

Introduction

- Statistics $\equiv$ Inference

Introduction

- Statistics $\equiv$ Inference
- How to extract information about scientific models from data.

Introduction

- Statistics $\equiv$ Inference
- 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.

Introduction

- Statistics $\equiv$ Inference
- 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.
- **Bayesians use Probability Distributions to quantify uncertainty.**

Probability distributions

Probability distributions

- As scientists, we are used to seeing error bars on results.

Probability distributions

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

Probability distributions

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

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

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:

$$m_1 = 36.4^{+5.5}_{-4.9}\, M_\odot, \qquad m_2 = 30.9^{+4.8}_{-4.4}\, M_\odot$$

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:

$$m_1 = 36.4^{+5.5}_{-4.9}\, M_\odot, \qquad m_2 = 30.9^{+4.8}_{-4.4}\, M_\odot$$

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

Probability distributions

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

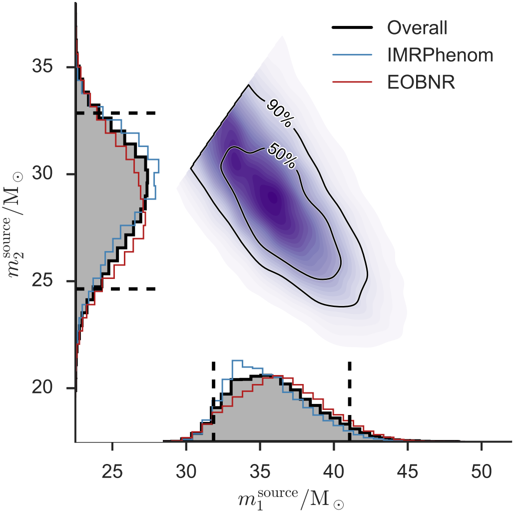
$$13.73 \pm 0.12 \text{ billion years old.}$$

- Masses of LIGO GW150914 binary merger:

$$m_1 = 36.4^{+5.5}_{-4.9}\, M_\odot, \qquad m_2 = 30.9^{+4.8}_{-4.4}\, M_\odot$$

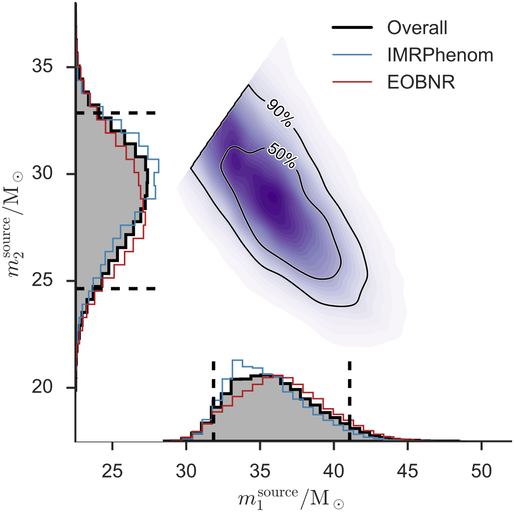
- These are called *credible intervals*, state that we are e.g. 90% confident of the value lying in this range.
- More importantly, these are *summary statistics*.

LIGO binary merger



$$m_1 = 36.4^{+5.5}_{-4.9}\ M_\odot,$$
$$m_2 = 30.9^{+4.8}_{-4.4}\ M_\odot$$

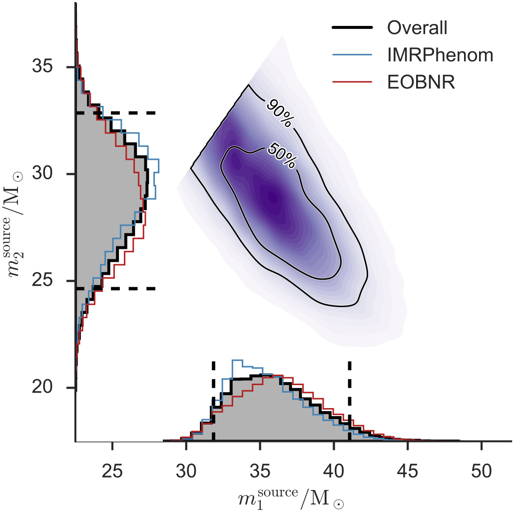
LIGO binary merger



$$m_1 = 36.4^{+5.5}_{-4.9}\ M_\odot,$$
$$m_2 = 30.9^{+4.8}_{-4.4}\ M_\odot$$

- Summary statistics summarise a full probability distribution.

LIGO binary merger



$$m_1 = 36.4^{+5.5}_{-4.9}\ M_\odot,$$
$$m_2 = 30.9^{+4.8}_{-4.4}\ M_\odot$$

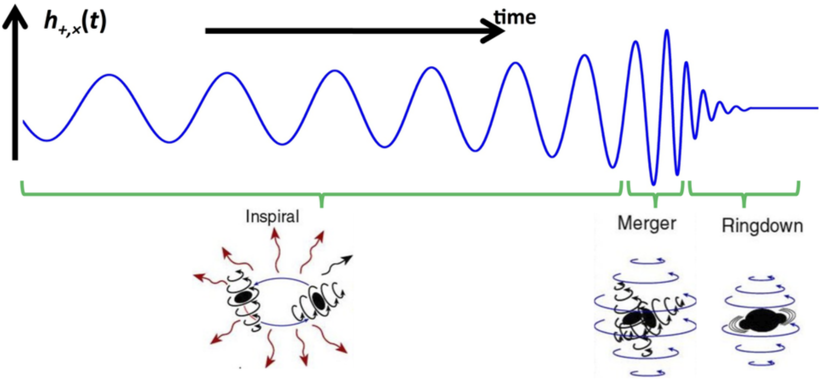
- Summary statistics summarise a full probability distribution.
- One goal of inference is to produce these probability distributions.

Extended example of inference: LIGO

- We will introduce the key concepts by discussing an extended example of the inference process.

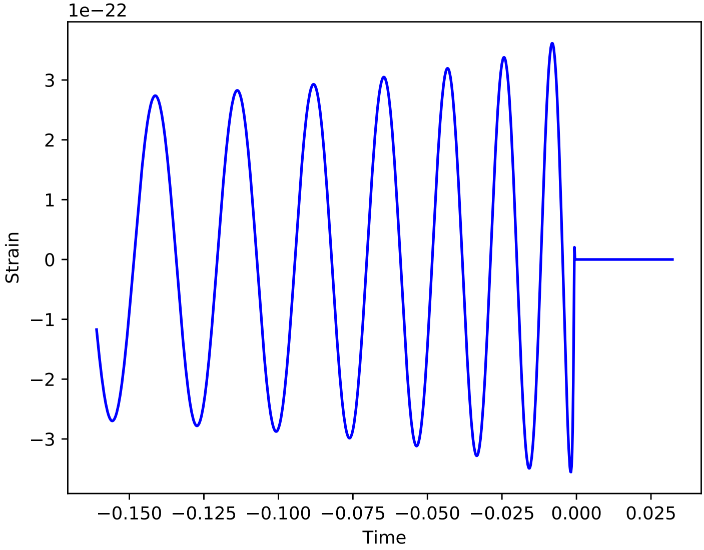
Theory

Extended example of inference: LIGO



The model M

Extended example of inference: LIGO



The parameters Θ of the model M

Extended example of inference: LIGO

Theoretical signal depends on:

The parameters Θ of the model M

Extended example of inference: LIGO

Theoretical signal depends on:

- m_1, m_2: mass of binary

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

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

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

The parameters Θ of the model M

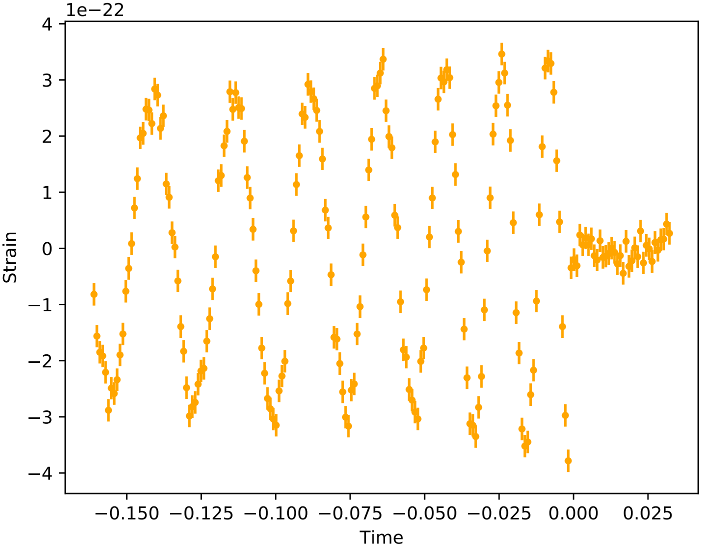
Extended example of inference: LIGO

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)

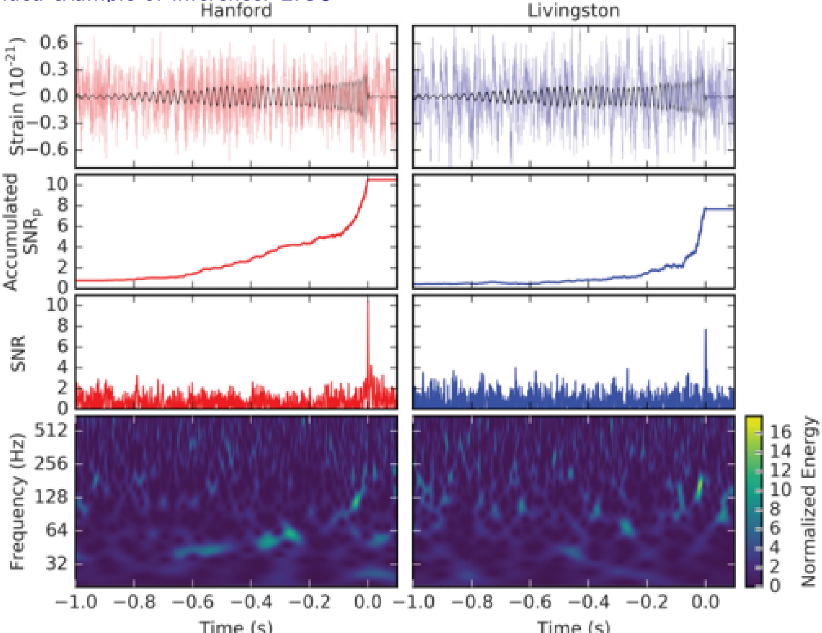
The data D

Extended example of inference: LIGO



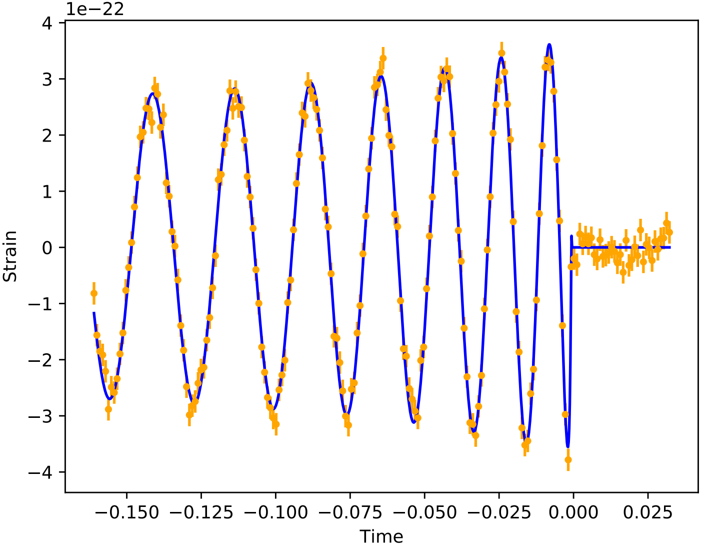
The data D

Extended example of inference: LIGO



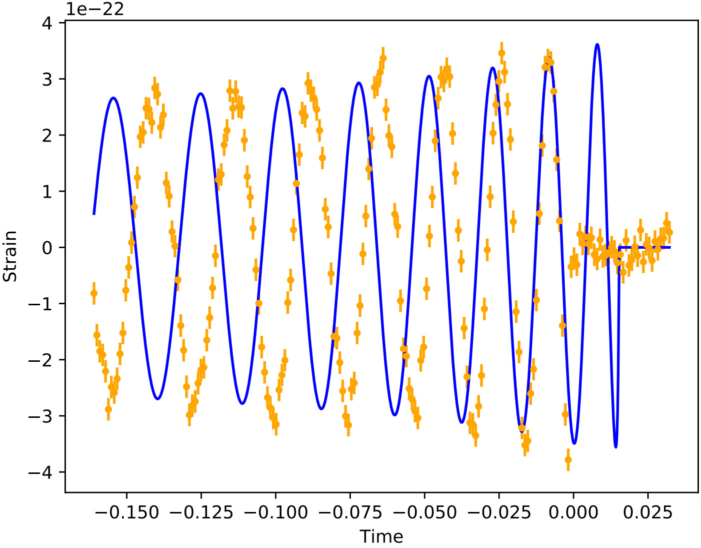
The Likelihood: well matched

Extended example of inference: LIGO



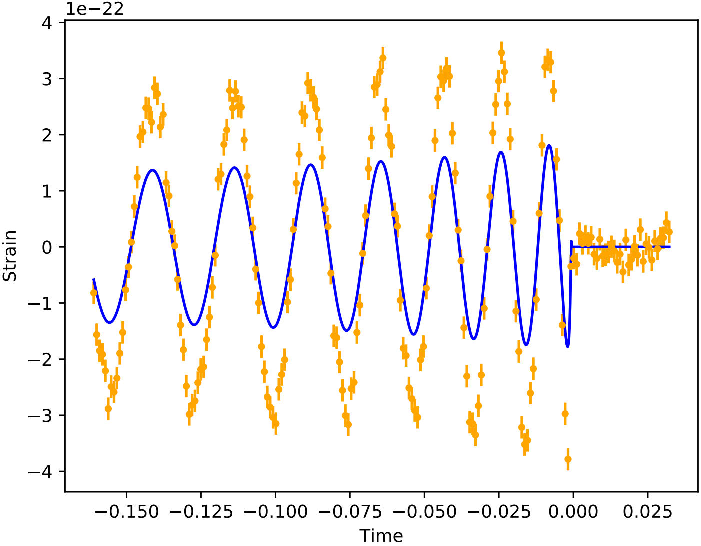
The Likelihood: coalescence off

Extended example of inference: LIGO



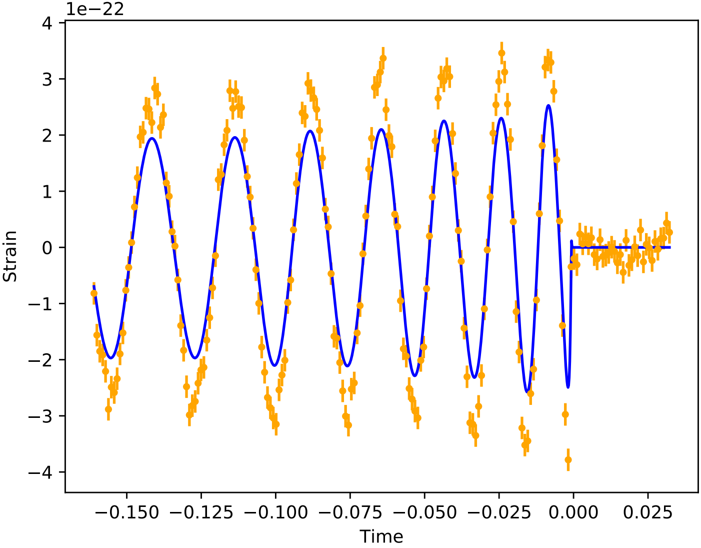
The Likelihood: too large luminosity distance

Extended example of inference: LIGO



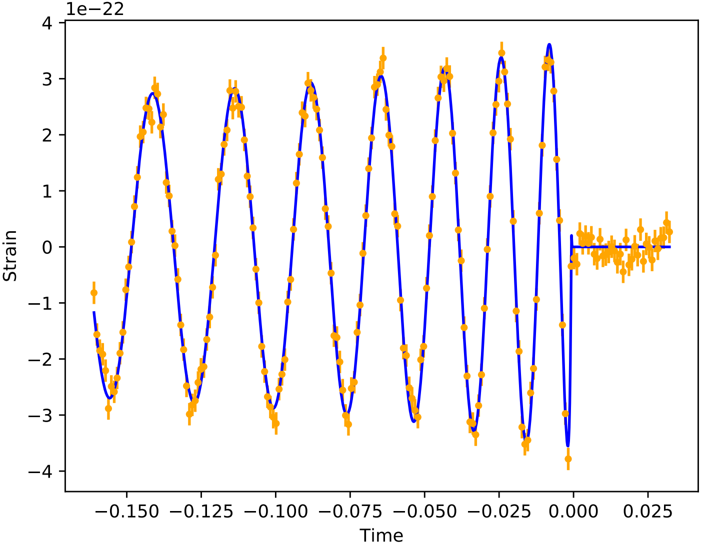
The Likelihood: incorrect inclination

Extended example of inference: LIGO



The Likelihood: 'Correct parameters'

Extended example of inference: LIGO



The likelihood $L(\Theta)$

- We assume that observed signal is combination of theoretical signal and Gaussian noise $O = S + N$.

The likelihood $L(\Theta)$

- We assume that observed signal is combination of theoretical signal and Gaussian noise $O = S + N$.
- Elementary probability tells us:

The likelihood $L(\Theta)$

- We assume that observed signal is combination of theoretical signal and Gaussian noise $O = S + N$.
- Elementary probability tells us:
  - $(t_i, h_i \pm \sigma_i)$: strain observed

The likelihood $L(\Theta)$

- We assume that observed signal is combination of theoretical signal and Gaussian noise $O = S + N$.
- Elementary probability tells us:
  - $(t_i, h_i \pm \sigma_i)$: strain observed
  - $h(t; \Theta)$: Theoretical strain

The likelihood $L(\Theta)$

- We assume that observed signal is combination of theoretical signal and Gaussian noise $O = S + N$.
- Elementary probability tells us:
  - $(t_i, h_i \pm \sigma_i)$: strain observed
  - $h(t; \Theta)$: Theoretical strain

$$P(h_i|\Theta, M) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{[h_i - h(t_i;\Theta)]^2}{2\sigma_i^2}\right)$$

The likelihood $L(\Theta)$

- We assume that observed signal is combination of theoretical signal and Gaussian noise $O = S + N$.
- Elementary probability tells us:
  - $(t_i, h_i \pm \sigma_i)$: strain observed
  - $h(t;\Theta)$: Theoretical strain

$$P(D|\Theta, M) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{[h_i - h(t_i;\Theta)]^2}{2\sigma_i^2}\right)$$

The likelihood $L(\Theta)$

- We assume that observed signal is combination of theoretical signal and Gaussian noise $O = S + N$.
- Elementary probability tells us:
  - $(t_i, h_i \pm \sigma_i)$: strain observed
  - $h(t;\Theta)$: Theoretical strain

$$L(\Theta) = P(D|\Theta, M) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{[h_i - h(t_i;\Theta)]^2}{2\sigma_i^2}\right)$$

The likelihood $L(\Theta)$

- We assume that observed signal is combination of theoretical signal and Gaussian noise $O = S + N$.
- Elementary probability tells us:
  - $(t_i, h_i \pm \sigma_i)$: strain observed
  - $h(t; \Theta)$: Theoretical strain

$$L(\Theta) = P(D|\Theta, M) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{[h_i - h(t_i; \Theta)]^2}{2\sigma_i^2}\right)$$

- We normally work with log-likelihoods, which turn $\prod \to \sum$.

Bayes' Theorem

Extended example of inference: LIGO

- Likelihood

Bayes' Theorem

Extended example of inference: LIGO

- Likelihood $\equiv$ Probability of data, given model parameters:

Bayes' Theorem

Extended example of inference: LIGO

- Likelihood $\equiv$ Probability of data, given model parameters:

$$L(\Theta) \equiv P(D|\Theta, M)$$

Bayes' Theorem

Extended example of inference: LIGO

- Likelihood $\equiv$ Probability of data, given model parameters:

$$L(\Theta) \equiv P(D|\Theta, M)$$

- What we want:
- *Posterior*

Bayes' Theorem

Extended example of inference: LIGO

- Likelihood $\equiv$ Probability of data, given model parameters:

$$L(\Theta) \equiv P(D|\Theta, M)$$

- What we want:
- *Posterior* $\equiv$ Probability of model parameters, given data:

Bayes' Theorem

Extended example of inference: LIGO

- Likelihood $\equiv$ Probability of data, given model parameters:

$$L(\Theta) \equiv P(D|\Theta, M)$$

- What we want:
- *Posterior* $\equiv$ Probability of model parameters, given data:

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

Bayes' Theorem

Extended example of inference: LIGO

- Likelihood $\equiv$ Probability of data, given model parameters:

$$L(\Theta) \equiv P(D|\Theta, M)$$

- What we want:
- *Posterior* $\equiv$ Probability of model parameters, given data:

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

- Use Bayes theorem:

Bayes' Theorem

Extended example of inference: LIGO

- Likelihood $\equiv$ Probability of data, given model parameters:

$$L(\Theta) \equiv P(D|\Theta, M)$$

- What we want:
- *Posterior* $\equiv$ Probability of model parameters, given data:

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

- Use Bayes theorem:

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

Bayes' Theorem

Extended example of inference: LIGO

- Likelihood $\equiv$ Probability of data, given model parameters:

$$L(\Theta) \equiv P(D|\Theta, M)$$

- What we want:
- *Posterior* $\equiv$ Probability of model parameters, given data:

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

- Use Bayes theorem:

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

$$\mathcal{P}(\Theta) = \frac{L(\Theta)\pi(\Theta)}{Z}$$

Bayes' Theorem

Extended example of inference: LIGO

- Likelihood $\equiv$ Probability of data, given model parameters:

$$L(\Theta) \equiv P(D|\Theta, M)$$

- What we want:
- *Posterior* $\equiv$ Probability of model parameters, given data:

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

- Use Bayes theorem:

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

$$\mathcal{P}(\Theta) = \frac{L(\Theta)\pi(\Theta)}{Z}$$

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

Prior $\pi(\Theta)$

Extended example of inference: LIGO

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

$$\mathcal{P}(\Theta) = \frac{L(\Theta)\pi(\Theta)}{Z}$$

Prior $\pi(\Theta)$

Extended example of inference: LIGO

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

$$\mathcal{P}(\Theta) = \frac{L(\Theta)\pi(\Theta)}{Z}$$

- Prior: $\pi(\Theta) \equiv P(\Theta|M)$.

Prior $\pi(\Theta)$

Extended example of inference: LIGO

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

$$\mathcal{P}(\Theta) = \frac{L(\Theta)\pi(\Theta)}{Z}$$

- Prior: $\pi(\Theta) \equiv P(\Theta|M)$.
- Represents initial degree of knowledge about the system:

Prior $\pi(\Theta)$

Extended example of inference: LIGO

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

$$\mathcal{P}(\Theta) = \frac{L(\Theta)\pi(\Theta)}{Z}$$

- Prior: $\pi(\Theta) \equiv P(\Theta|M)$.
- Represents initial degree of knowledge about the system:
- e.g. $\pi(\theta,\phi)\, d\theta d\phi = -\frac{1}{2}\sin\theta\, d\theta d\phi$

Prior $\pi(\Theta)$

Extended example of inference: LIGO

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

$$\mathcal{P}(\Theta) = \frac{L(\Theta)\pi(\Theta)}{Z}$$

- Prior: $\pi(\Theta) \equiv P(\Theta|M)$.
- Represents initial degree of knowledge about the system:
- e.g. $\pi(\theta, \phi)\, d\theta d\phi = -\frac{1}{2}\sin\theta\, d\theta d\phi$
- Most Bayesian approaches are sensitive to this, and rightly so.

Evidence Z

Extended example of inference: LIGO

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

$$\mathcal{P}(\Theta) = \frac{L(\Theta)\pi(\Theta)}{Z}$$

- Evidence: $Z \equiv P(D|M)$

Evidence Z

Extended example of inference: LIGO

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

$$\mathcal{P}(\Theta) = \frac{L(\Theta)\pi(\Theta)}{Z}$$

- Evidence: $Z \equiv P(D|M) = \int P(D|\Theta,M)P(\Theta|M)d\Theta$.

Evidence Z

Extended example of inference: LIGO

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

$$\mathcal{P}(\Theta) = \frac{L(\Theta)\pi(\Theta)}{Z}$$

- Evidence: $Z \equiv P(D|M) = \int L(\Theta)\pi(\Theta)d\Theta$.

Evidence Z

Extended example of inference: LIGO

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

$$\mathcal{P}(\Theta) = \frac{L(\Theta)\pi(\Theta)}{Z}$$

- Evidence: $Z \equiv P(D|M) = \int L(\Theta)\pi(\Theta)d\Theta$.
- Normalising constant.

Evidence Z

Extended example of inference: LIGO

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

$$\mathcal{P}(\Theta) = \frac{L(\Theta)\pi(\Theta)}{Z}$$

- Evidence: $Z \equiv P(D|M) = \int L(\Theta)\pi(\Theta)d\Theta$.
- Normalising constant.
- Difficult to compute.

Evidence Z

Extended example of inference: LIGO

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

- Evidence: $Z \equiv P(D|M) = \int L(\Theta)\pi(\Theta)d\Theta$.
- Normalising constant.
- Difficult to compute.
- Still extremely important.

Posterior $\mathcal{P}$

Extended example of inference: LIGO

- Cannot plot the full posterior distribution:

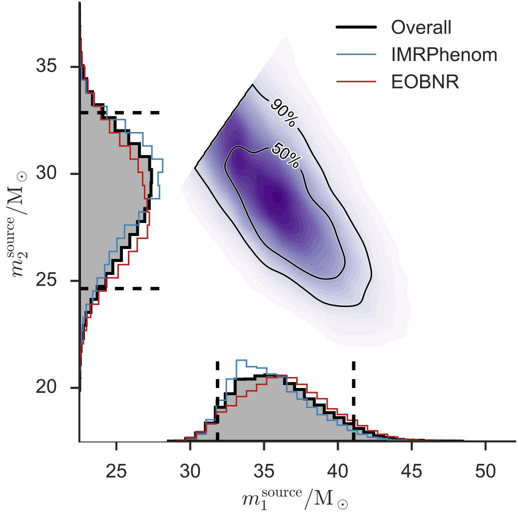
$$\mathcal{P}(\Theta) \equiv P(m_1, m_2, \theta, \phi, r, \Phi_c, t_c, i, \theta_{\mathrm{sky}}|D, M)$$

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

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

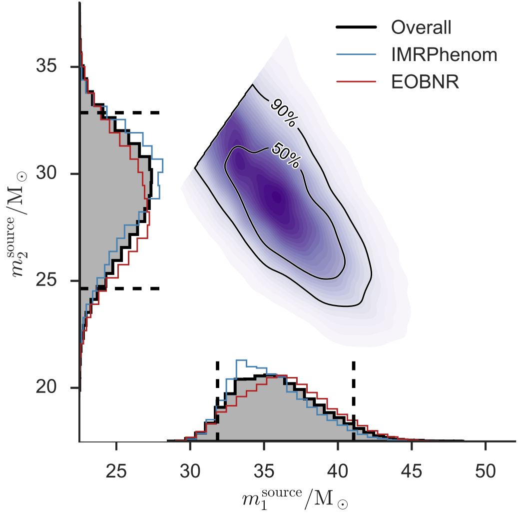
Posterior $\mathcal{P}$

Extended example of inference: LIGO



Posterior $\mathcal{P}$

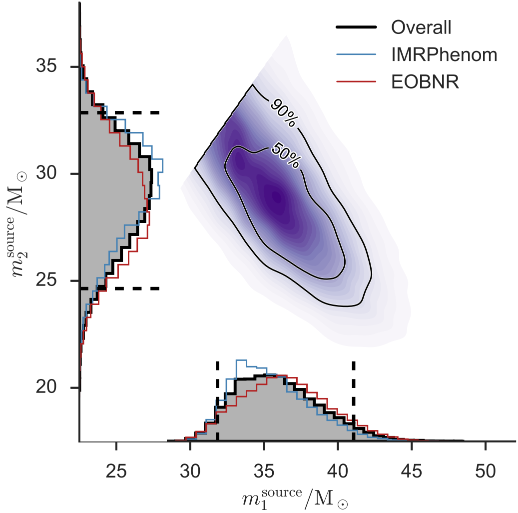
Extended example of inference: LIGO



- May do this for each pair of parameters

Posterior $\mathcal{P}$

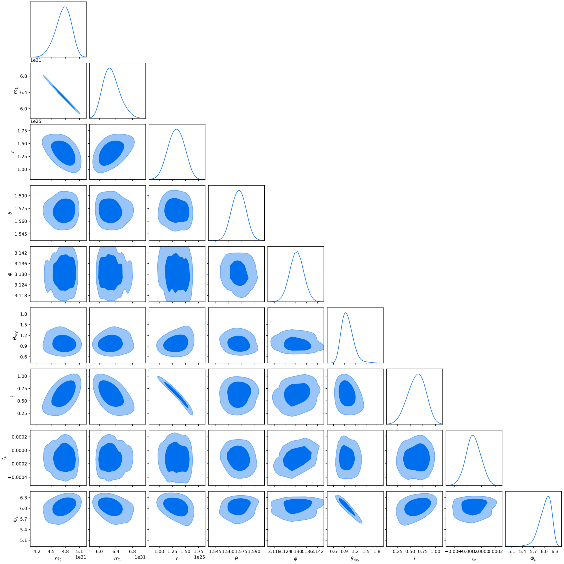
Extended example of inference: LIGO



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

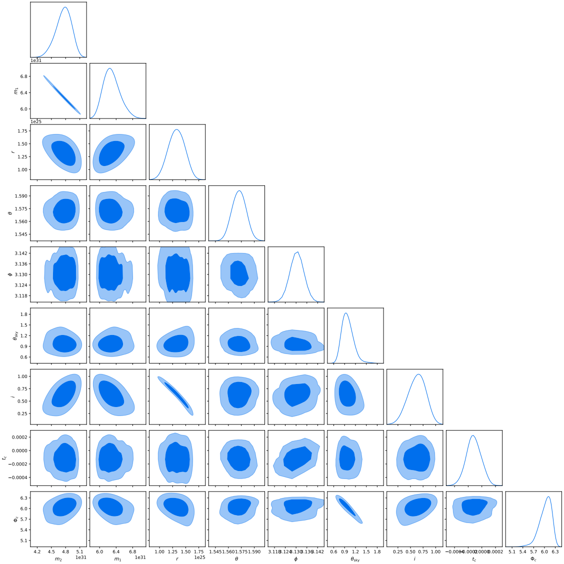
Posterior $\mathcal{P}$

Extended example of inference: LIGO



Posterior $\mathcal{P}$

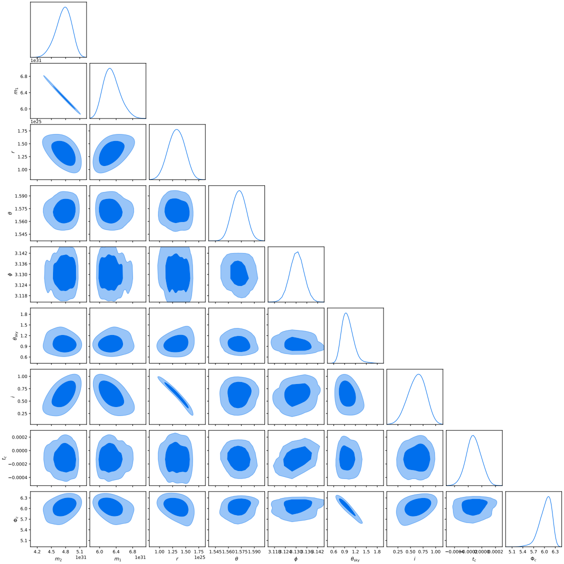
Extended example of inference: LIGO



- Does give insight

Posterior $\mathcal{P}$

Extended example of inference: LIGO



- Does give insight
- Not the full picture

Evidences and model comparison

Extended example of inference: LIGO

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.

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.

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".

Bayes' Theorem (again)

Bayes' Theorem (again)

What does data tell us about our model M_i?

Bayes' Theorem (again)

What does data tell us about our model M_i?

$$\mathrm{P}(M_i|D) = \frac{\mathrm{P}(D|M_i)\mathrm{P}(M_i)}{\mathrm{P}(D)}$$

Bayes' Theorem (again)

What does data tell us about our model M_i?

$$\mathrm{P}(M_i|D) = \frac{\mathrm{P}(D|M_i)\mathrm{P}(M_i)}{\mathrm{P}(D)}$$

$$\mathrm{P}(M_i|D) = \frac{Z_i\;\mu_i}{\sum_k Z_k\;\mu_k}$$

Bayes' Theorem (again)

What does data tell us about our model M_i?

$$\mathrm{P}(M_i|D) = \frac{\mathrm{P}(D|M_i)\mathrm{P}(M_i)}{\mathrm{P}(D)}$$

$$\mathrm{P}(M_i|D) = \frac{Z_i\ \mu_i}{\sum_k Z_k\ \mu_k}$$

- Evidences allow us to determine what weighting we should give to each model in light of the data.

Bayes' Theorem (again)

What does data tell us about our model M_i?

$$\mathrm{P}(M_i|D) = \frac{\mathrm{P}(D|M_i)\mathrm{P}(M_i)}{\mathrm{P}(D)}$$

$$\mathrm{P}(M_i|D) = \frac{Z_i\ \mu_i}{\sum_k Z_k\ \mu_k}$$

- Evidences allow us to determine what weighting we should give to each model in light of the data.
- e.g. EOBNR (Effective One Body) vs IMRPhenom (Inspiral-merger-ringdown).

Bayes' Theorem (again)

What does data tell us about our model M_i?

$$\mathrm{P}(M_i|D) = \frac{\mathrm{P}(D|M_i)\mathrm{P}(M_i)}{\mathrm{P}(D)}$$

$$\mathrm{P}(M_i|D) = \frac{Z_i\ \mu_i}{\sum_k Z_k\ \mu_k}$$

- Evidences allow us to determine what weighting we should give to each model in light of the data.
- e.g. EOBNR (Effective One Body) vs IMRPhenom (Inspiral-merger-ringdown).
- Or testing modified theories of gravity.

Bayes' Theorem (again)

What does data tell us about our model M_i?

$$\mathrm{P}(M_i|D) = \frac{\mathrm{P}(D|M_i)\mathrm{P}(M_i)}{\mathrm{P}(D)}$$

$$\mathrm{P}(M_i|D) = \frac{Z_i\ \mu_i}{\sum_k Z_k\ \mu_k}$$

- Evidences allow us to determine what weighting we should give to each model in light of the data.
- e.g. EOBNR (Effective One Body) vs IMRPhenom (Inspiral-merger-ringdown).
- Or testing modified theories of gravity.

Model averaging:

- Multiple models with posterior on the same parameter: $\mathrm{P}(y|M_i, D)$

$$\mathrm{P}(y|D) = \sum_i \mathrm{P}(y|M_i, D)\mathrm{P}(M_i|D)$$

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.

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

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

Parameter estimation

Another example.

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

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

Parameter estimation

Another example.

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

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

Parameter estimation

Another example.

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

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

Parameter estimation

Another example.

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

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

Parameter estimation

Another example.

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

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

Parameter estimation

Another example.

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

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

Parameter estimation

Another example.

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

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

Parameter estimation

Another example.

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

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

Parameter estimation

Another example.

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

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

Parameter estimation

Another example.

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

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

Parameter estimation

Another example.

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

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

Parameter estimation

Another example.

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

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

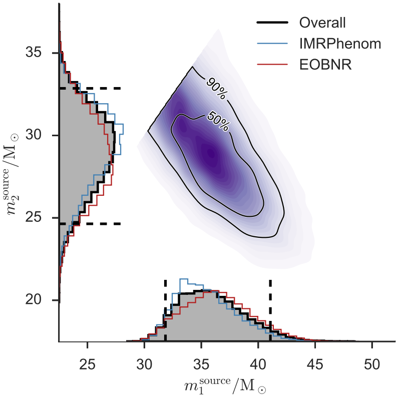
- Parameter estimation: $L, \pi \to \mathcal{P}$: model parameters
- Model comparison: $L, \pi \to Z$: how good model is

Sampling

How to describe a high-dimensional posterior

Sampling

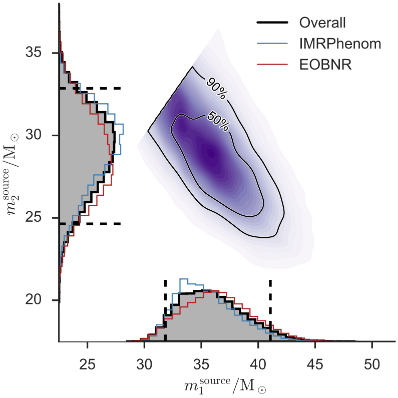
How to describe a high-dimensional posterior



Sampling

How to describe a high-dimensional posterior

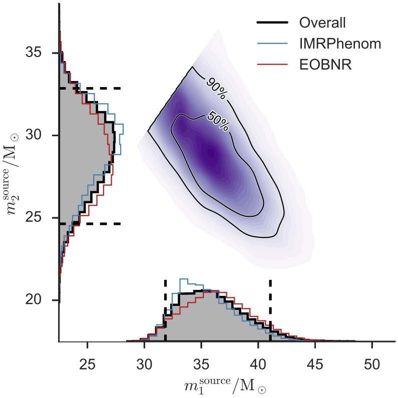
- In high dimensions, posterior $\mathcal{P}$ occupies a vanishingly small region of the prior π.



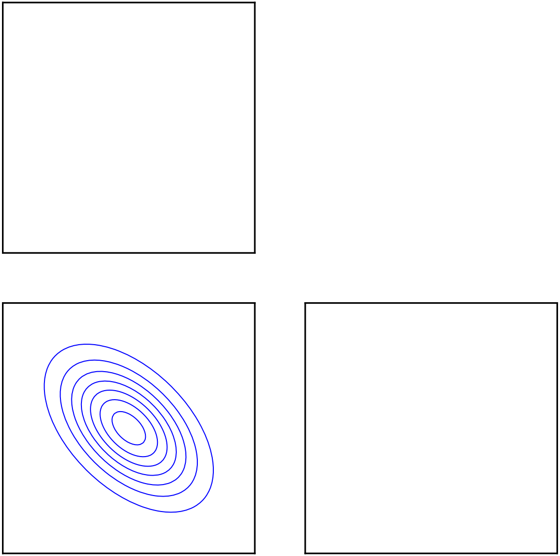
Sampling

How to describe a high-dimensional posterior

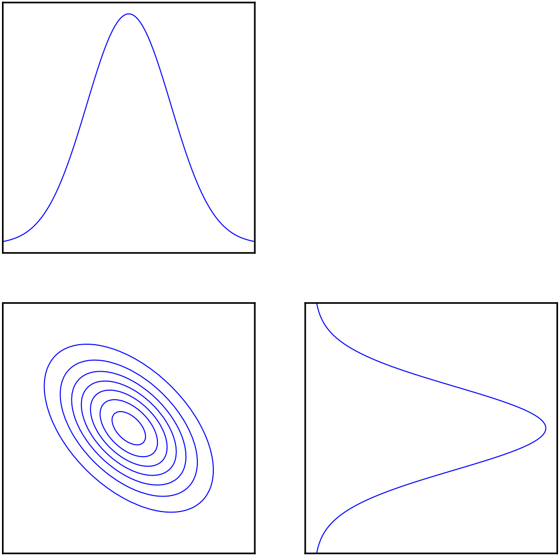
- In high dimensions, posterior $\mathcal{P}$ occupies a vanishingly small region of the prior π.
- *Sampling* the posterior is an excellent compression scheme.



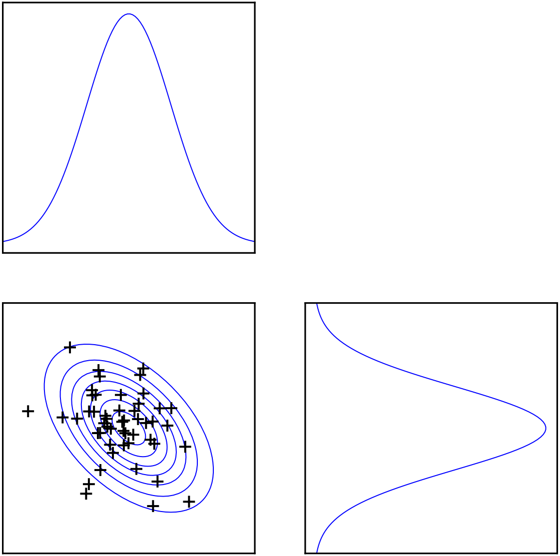
Sampling



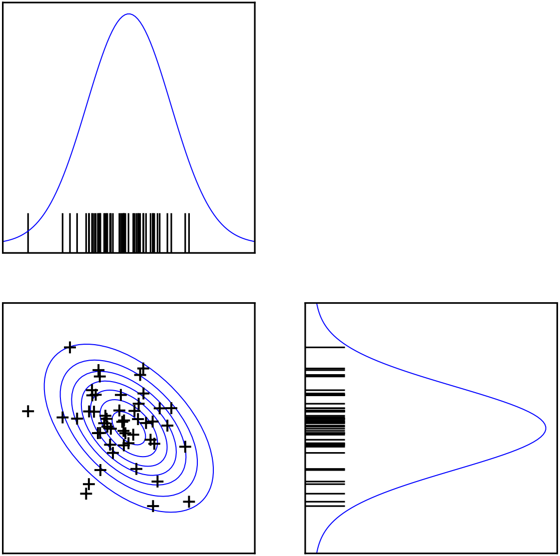
Sampling



Sampling



Sampling



Why do sampling?

Marginalisation over the posterior

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

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:

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

Why do sampling?

Marginalisation over the posterior

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

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

Why do sampling?

Marginalisation over the posterior

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

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:

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

- 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

Parameter estimation

- The name of the game is therefore drawing samples S from the posterior $\mathcal{P}$ with the minimum number of likelihood calls.

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.

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

Current sampling approaches

1. Metropolis Hastings.

Current sampling approaches

1. Metropolis Hastings.
2. Hamiltonian Monte-Carlo (HMC).

Current sampling approaches

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

Metropolis Hastings

Metropolis Hastings

- Turn the N-dimensional problem into a one-dimensional one.

Metropolis Hastings

- Turn the N-dimensional problem into a one-dimensional one.
    1. Pick random direction

Metropolis Hastings

- Turn the N-dimensional problem into a one-dimensional one.
  1. Pick random direction
  2. Choose step length

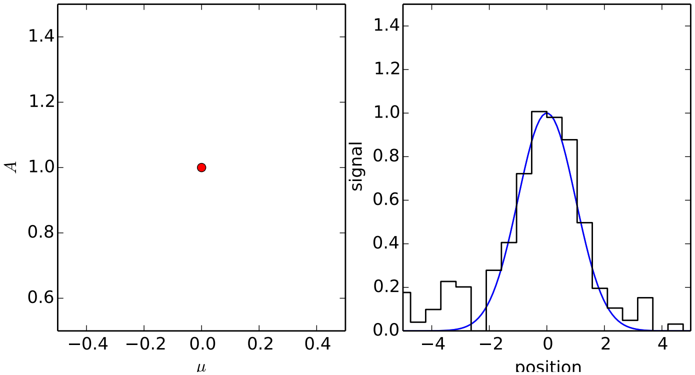
Metropolis Hastings

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

Metropolis Hastings

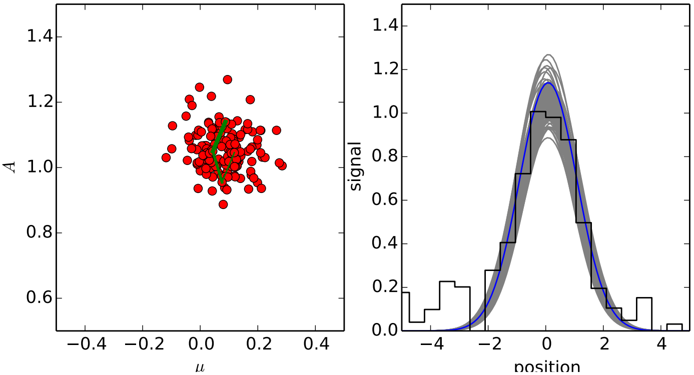
- Turn the N-dimensional problem into a one-dimensional one.
  1. Pick random direction
  2. Choose step length
  3. If uphill, make step...
  4. ...otherwise sometimes make step.

Metropolis Hastings



Metropolis Hastings

Metropolis Hastings



Metropolis Hastings

Struggles with. . .

Metropolis Hastings

Struggles with...

1. Burn in

Metropolis Hastings

Struggles with...

1. Burn in
2. Multimodality

Metropolis Hastings

Struggles with...

1. Burn in
2. Multimodality
3. Correlated Peaks

Metropolis Hastings

Struggles with...

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

Hamiltonian Monte-Carlo

Hamiltonian Monte-Carlo

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

Hamiltonian Monte-Carlo

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

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

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

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.

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

Ensemble sampling

- Instead of one walker, evolve a set of n walkers.

Ensemble sampling

- Instead of one walker, evolve a set of n walkers.
- Can use information present in ensemble to guide proposals.

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.

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

The fundamental issue with all of the above

- They don't give you evidences!

$$\mathcal{Z} = \mathrm{P}(D|M)$$

$$= \langle \mathcal{L} \rangle_\pi$$

- 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

The fundamental issue with all of the above

- They don't give you evidences!

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

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.

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.

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.

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.

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?

What is nested sampling?

- Nested sampling is an alternative way of sampling posteriors.

What is nested sampling?

- Nested sampling is an alternative way of sampling posteriors.
- Uses ensemble sampling to compress prior to posterior.

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!

Nested Sampling

John Skilling's alternative to traditional MCMC!

New procedure:

Nested Sampling

John Skilling's alternative to traditional MCMC!

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

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).

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

Nested Sampling

John Skilling's alternative to traditional MCMC!

New procedure:

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

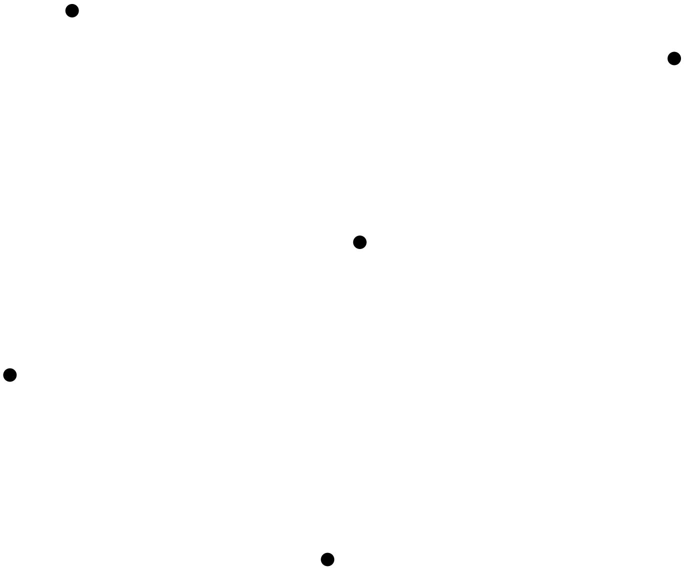
S_0: Generate n samples uniformly over the space (from the prior π).

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*.

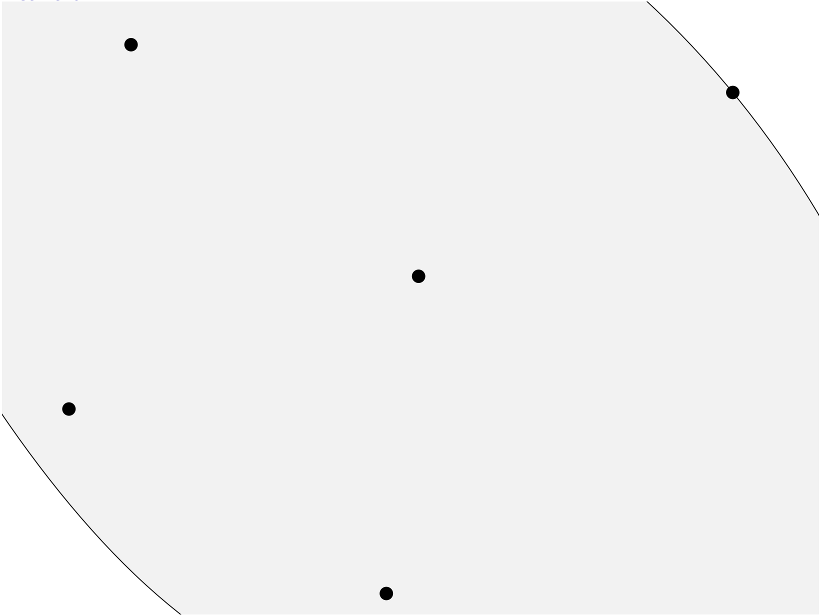
Nested Sampling

Graphical aid



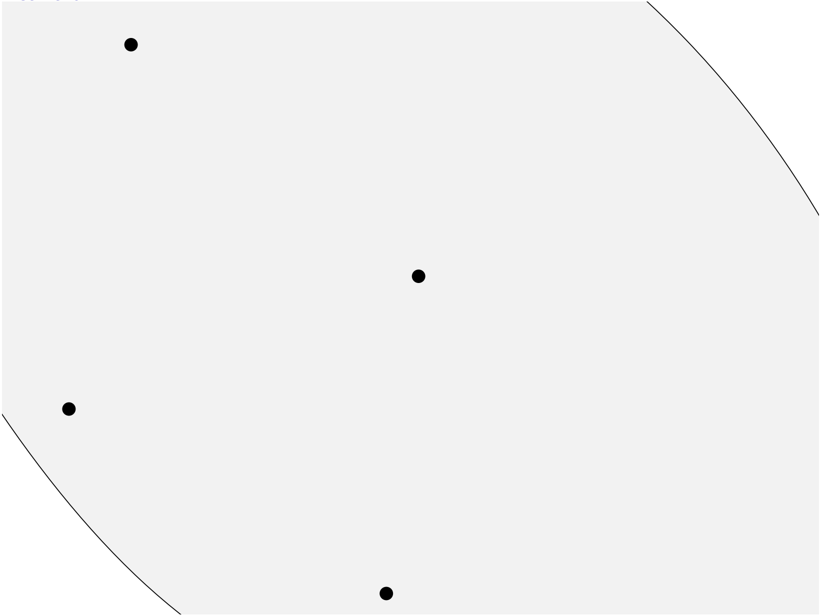
Nested Sampling

Graphical aid



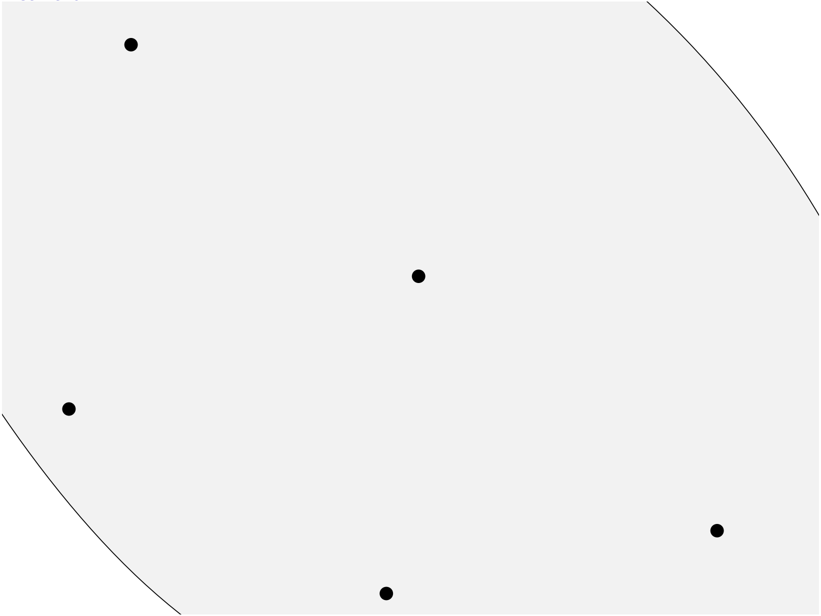
Nested Sampling

Graphical aid



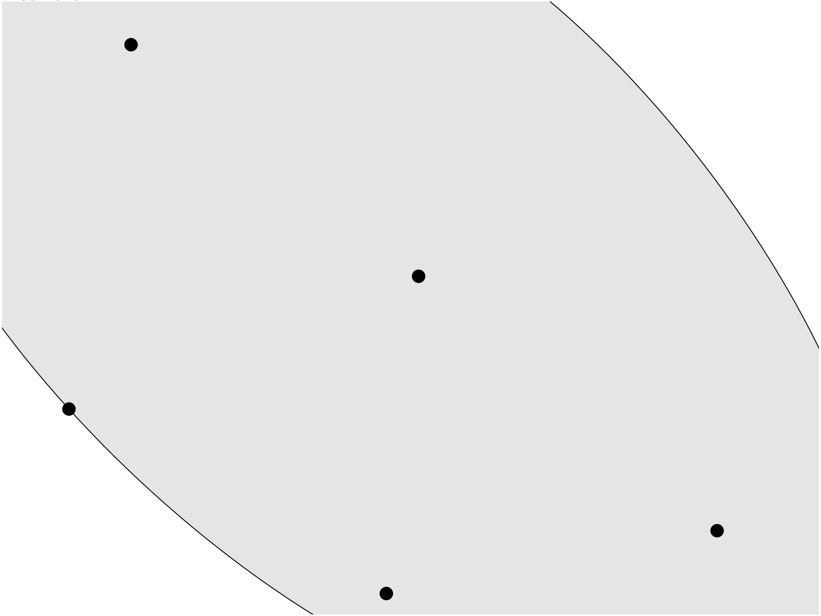
Nested Sampling

Graphical aid



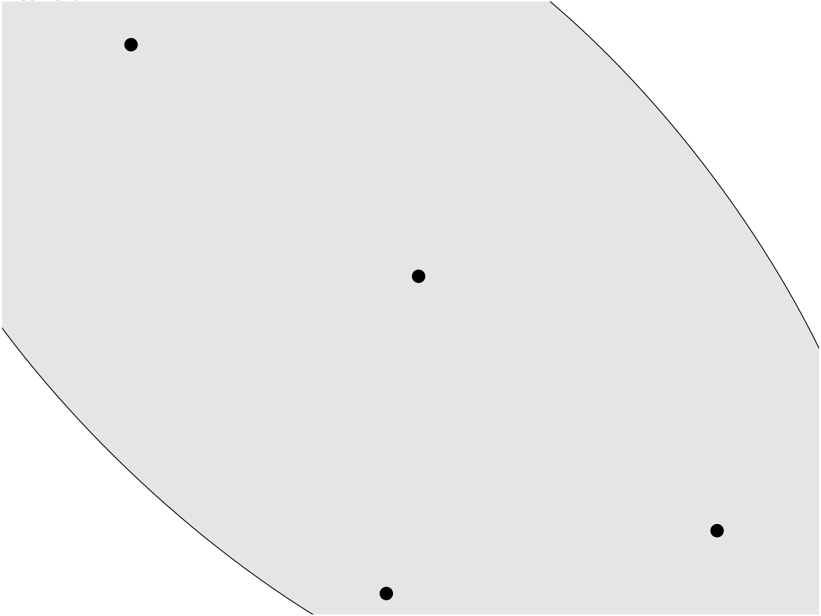
Nested Sampling

Graphical aid



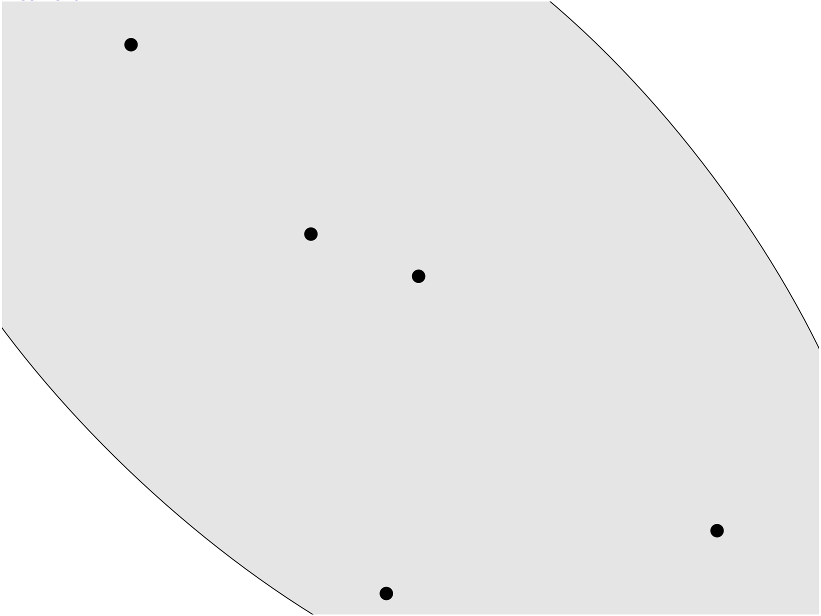
Nested Sampling

Graphical aid



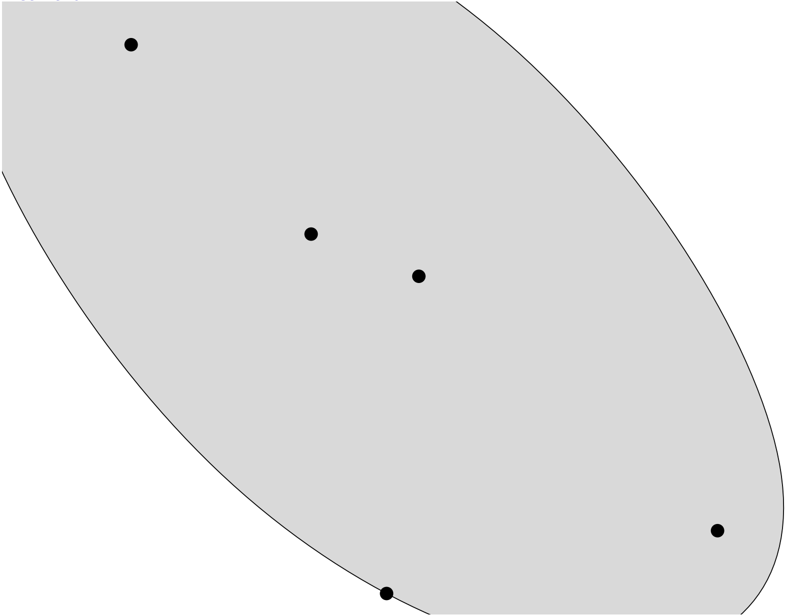
Nested Sampling

Graphical aid



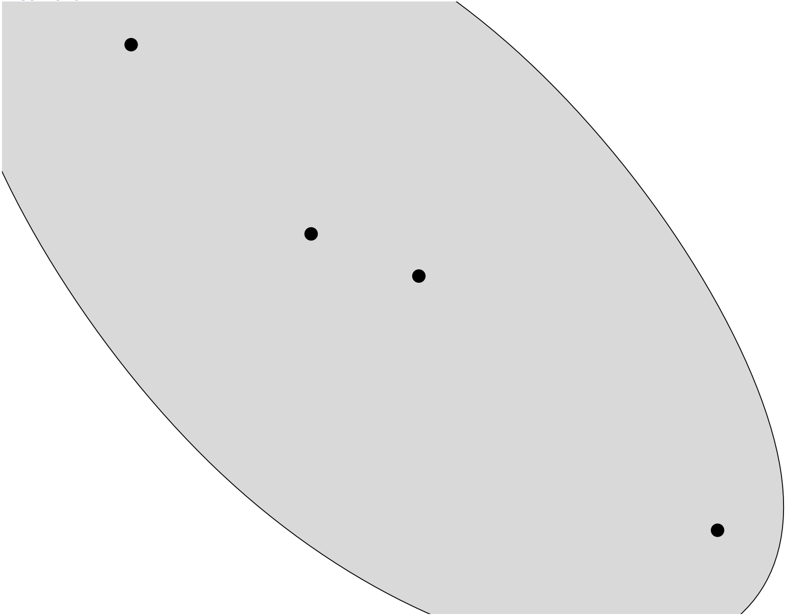
Nested Sampling

Graphical aid



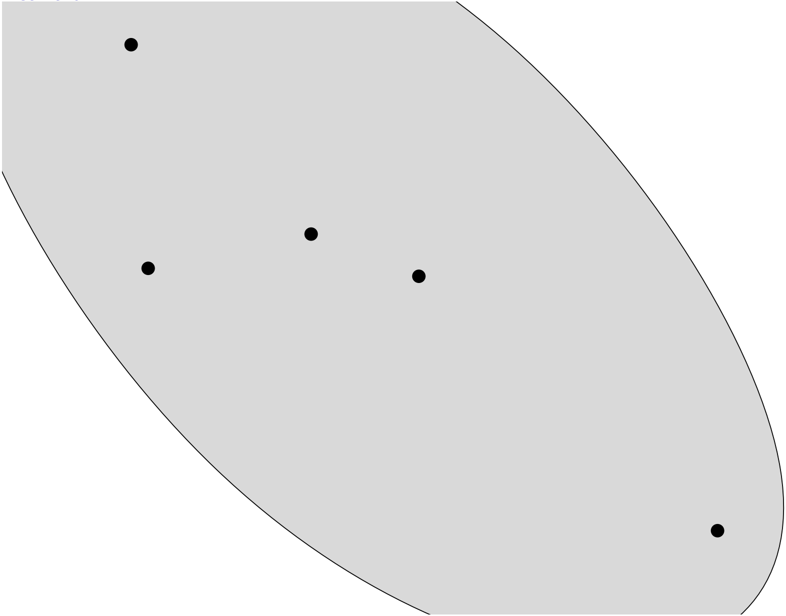
Nested Sampling

Graphical aid



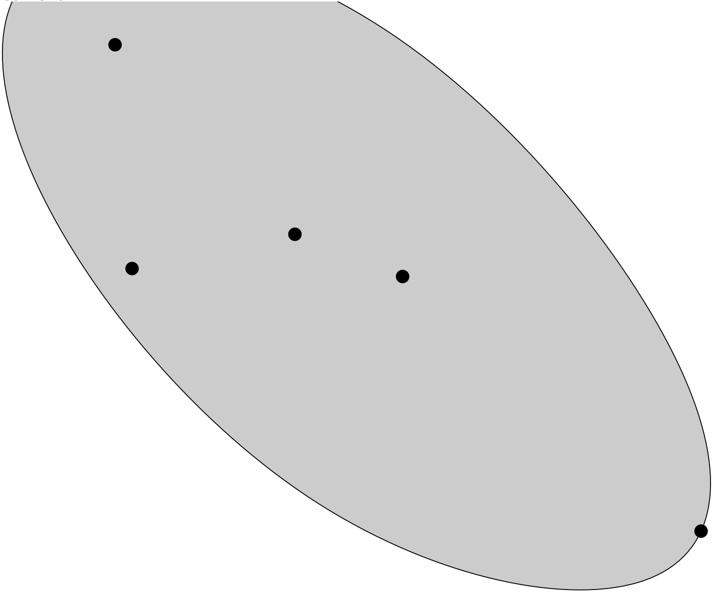
Nested Sampling

Graphical aid



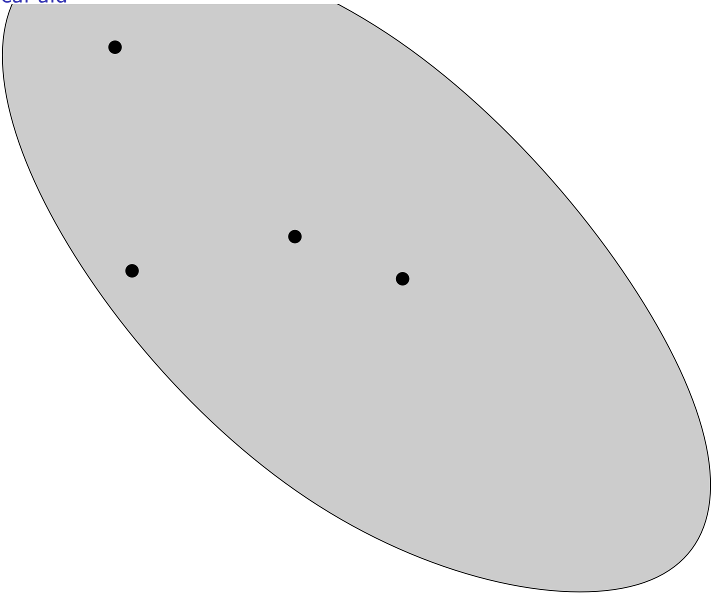
Nested Sampling

Graphical aid



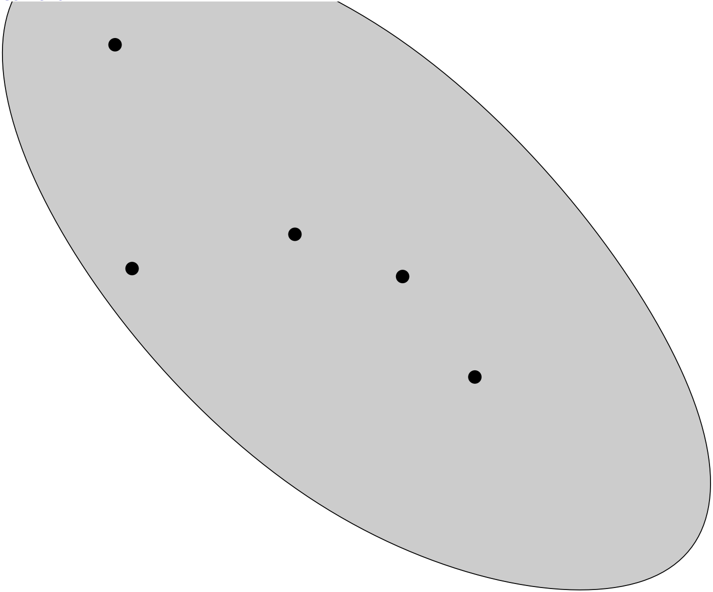
Nested Sampling

Graphical aid



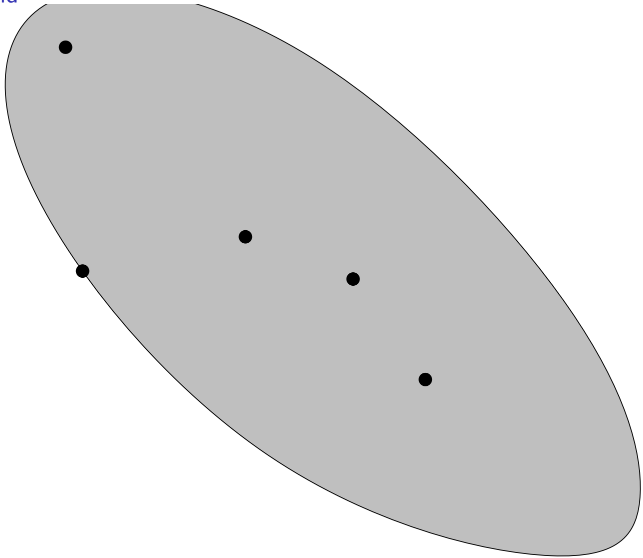
Nested Sampling

Graphical aid



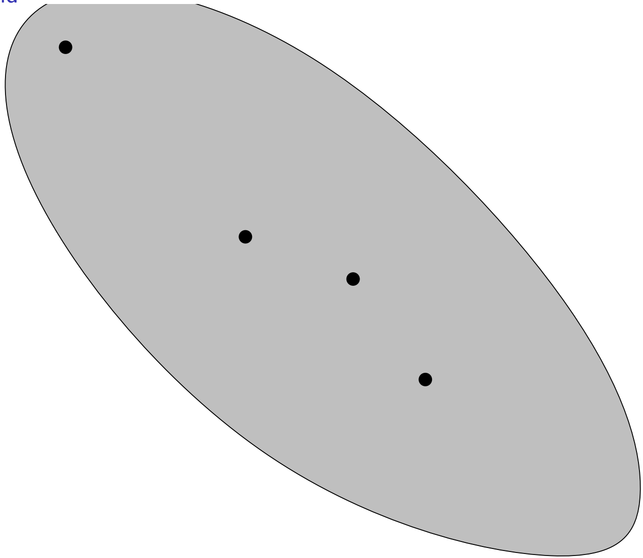
Nested Sampling

Graphical aid



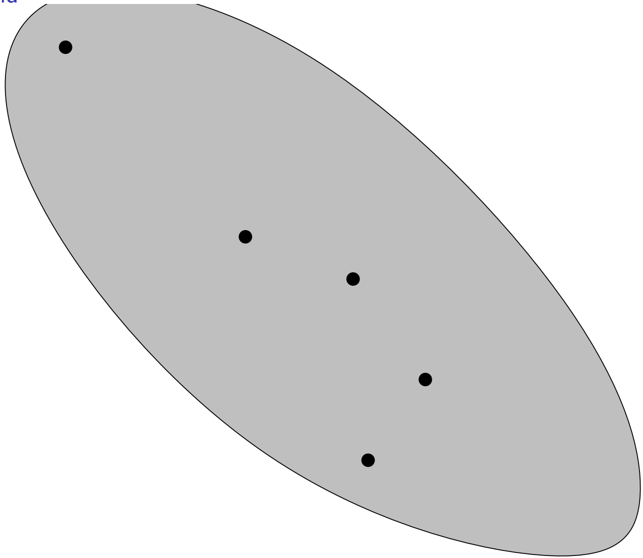
Nested Sampling

Graphical aid



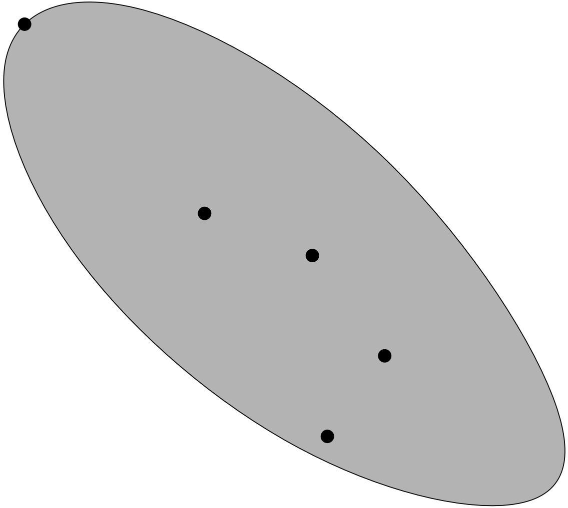
Nested Sampling

Graphical aid



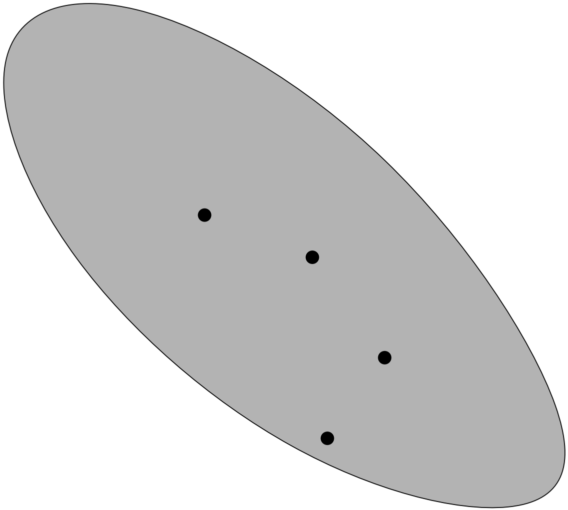
Nested Sampling

Graphical aid



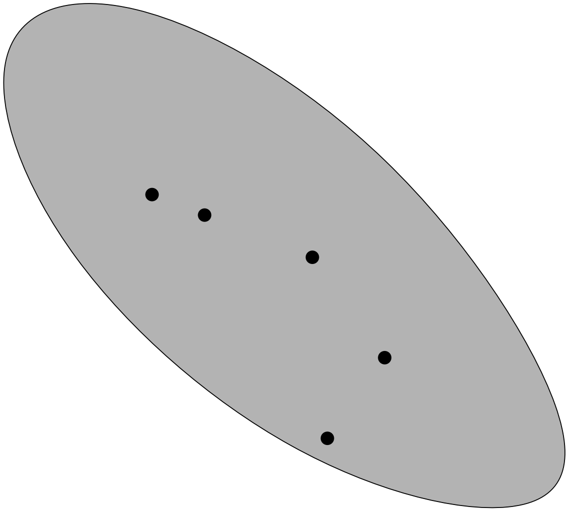
Nested Sampling

Graphical aid



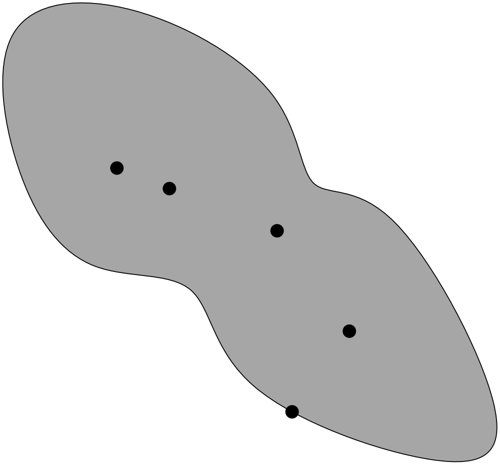
Nested Sampling

Graphical aid



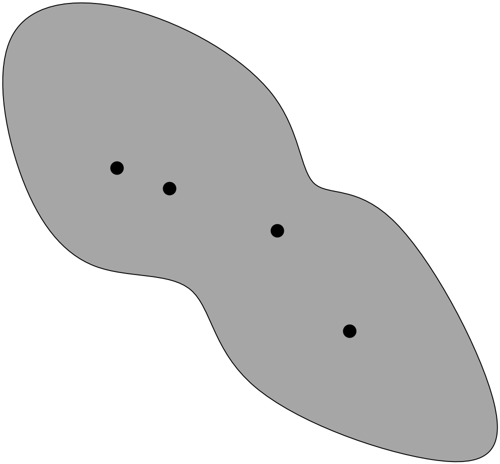
Nested Sampling

Graphical aid



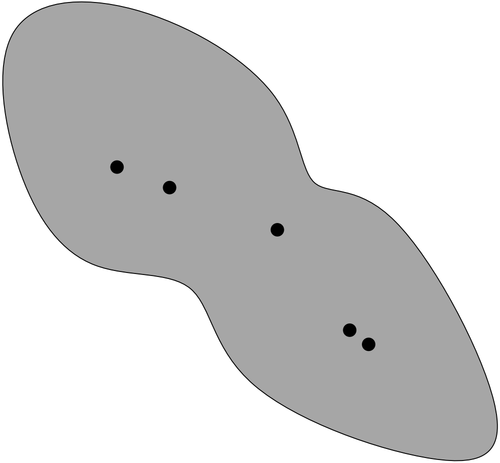
Nested Sampling

Graphical aid



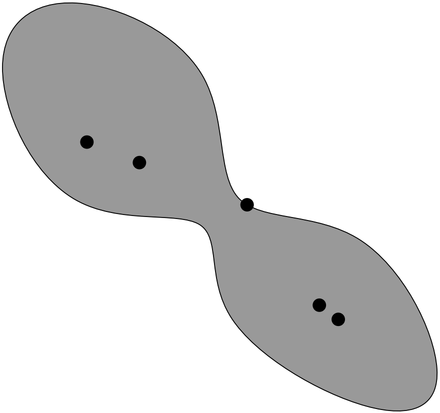
Nested Sampling

Graphical aid



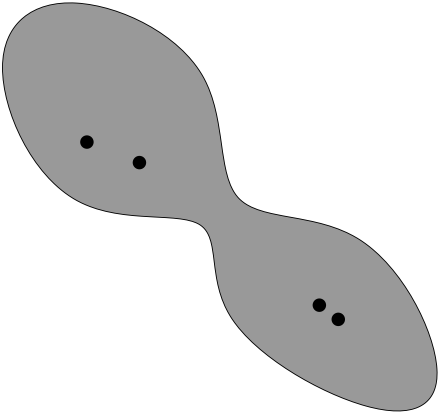
Nested Sampling

Graphical aid



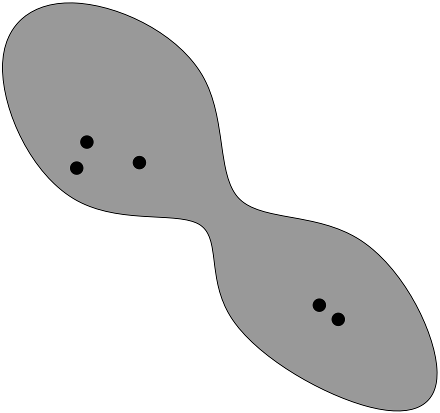
Nested Sampling

Graphical aid



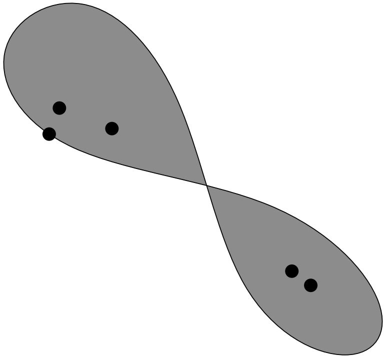
Nested Sampling

Graphical aid



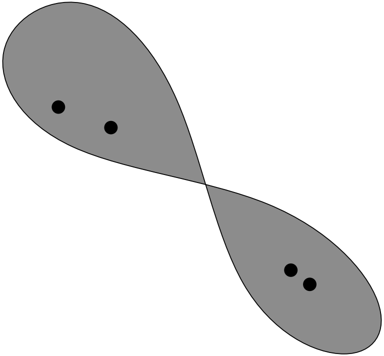
Nested Sampling

Graphical aid



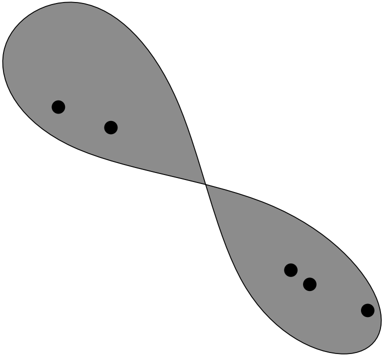
Nested Sampling

Graphical aid



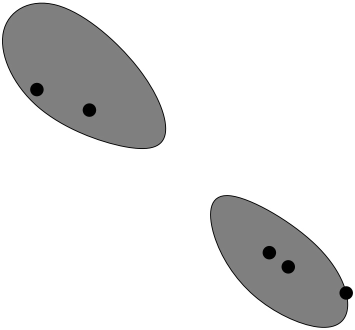
Nested Sampling

Graphical aid



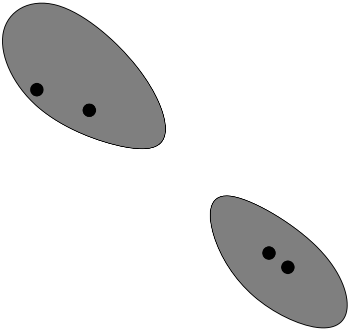
Nested Sampling

Graphical aid



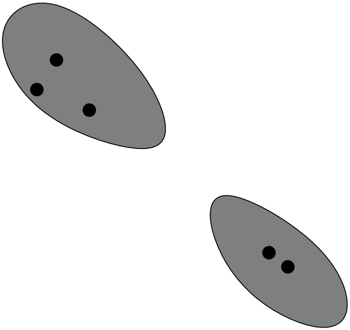
Nested Sampling

Graphical aid



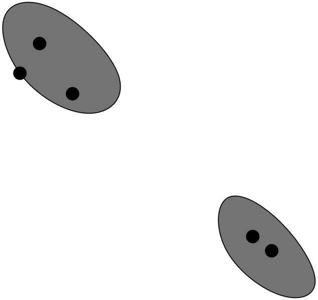
Nested Sampling

Graphical aid



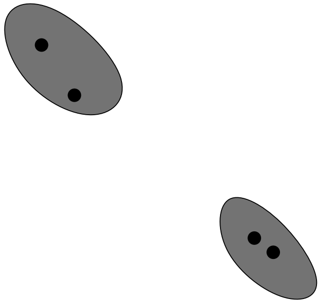
Nested Sampling

Graphical aid



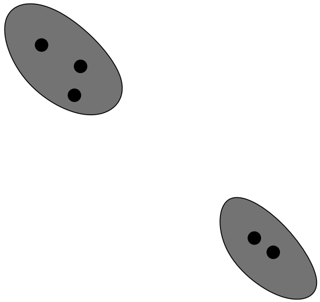
Nested Sampling

Graphical aid



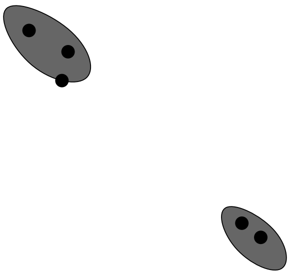
Nested Sampling

Graphical aid



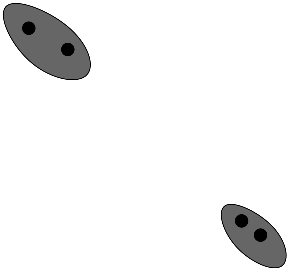
Nested Sampling

Graphical aid



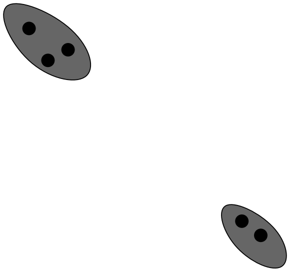
Nested Sampling

Graphical aid



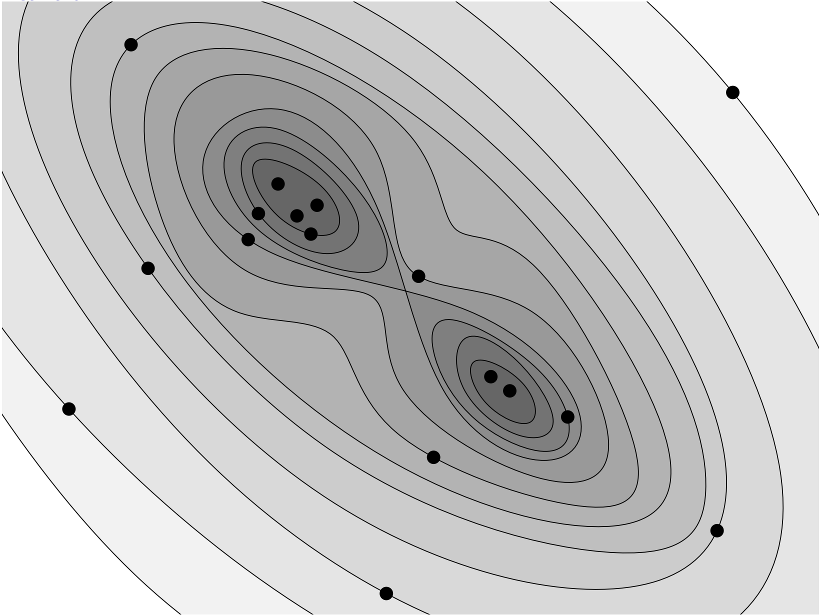
Nested Sampling

Graphical aid



Nested Sampling

Graphical aid



Nested sampling

Nested sampling

- The set of dead points are posterior samples with an appropriate weighting factor

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

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*

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

Sampling within an iso-likelihood contour

Previous attempts

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

Sampling within an iso-likelihood contour

Previous attempts

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

- Suffers in high dimensions

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)

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)

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

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).

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

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.

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).

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”.

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).

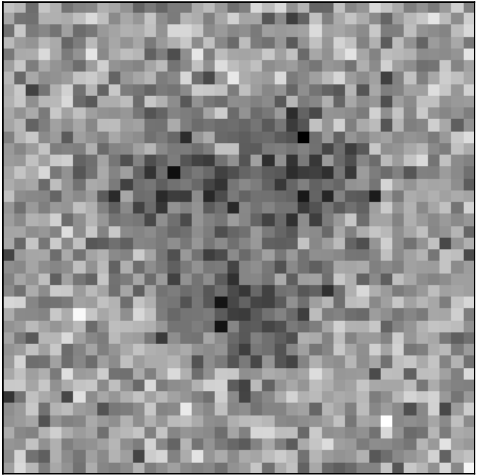
- Very promising
- Still needs tuning.

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

- Current “state-of-the-art”.
- PolyChord 2.0 imminent.

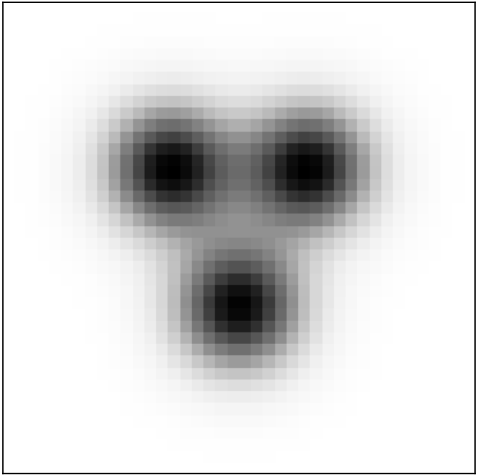
Object detection

Toy problem



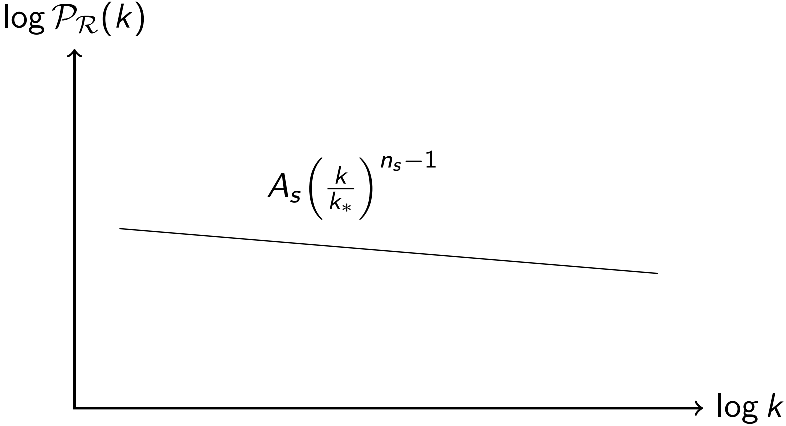
Object detection

Toy problem



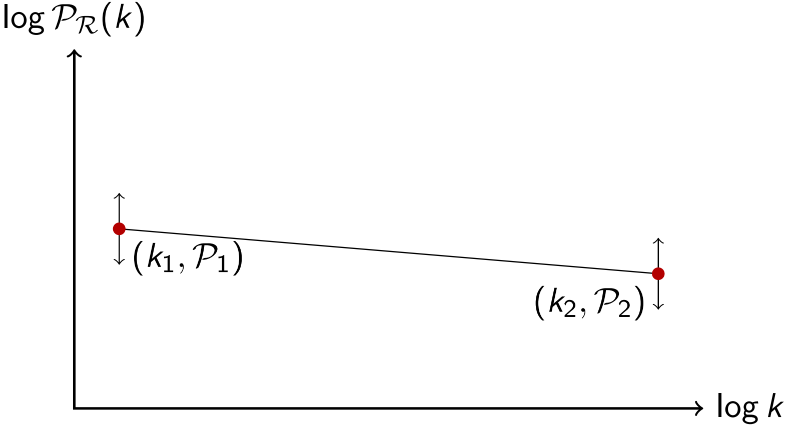
PolyChord in action

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



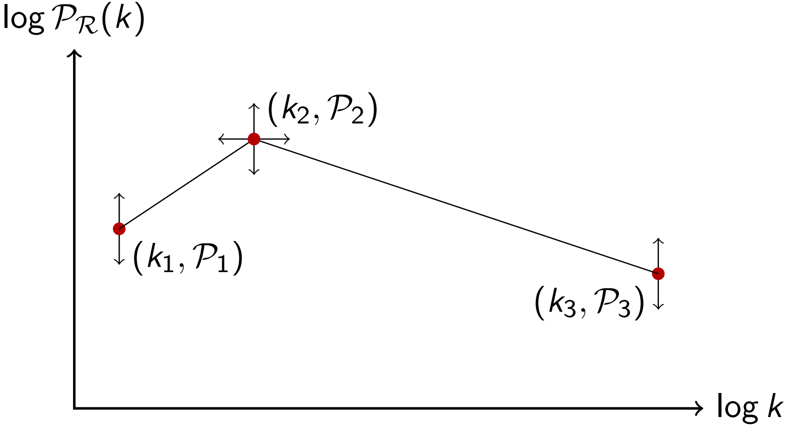
PolyChord in action

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



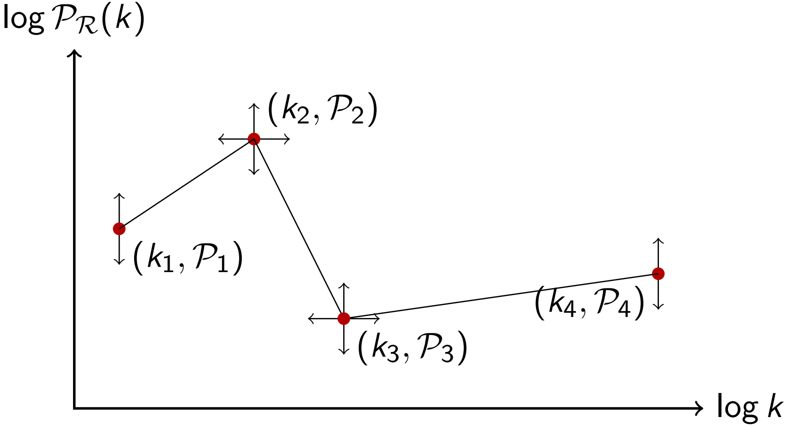
PolyChord in action

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



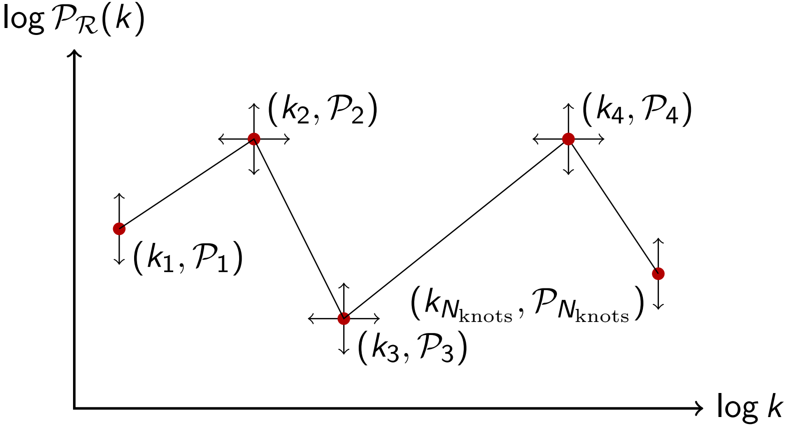
PolyChord in action

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



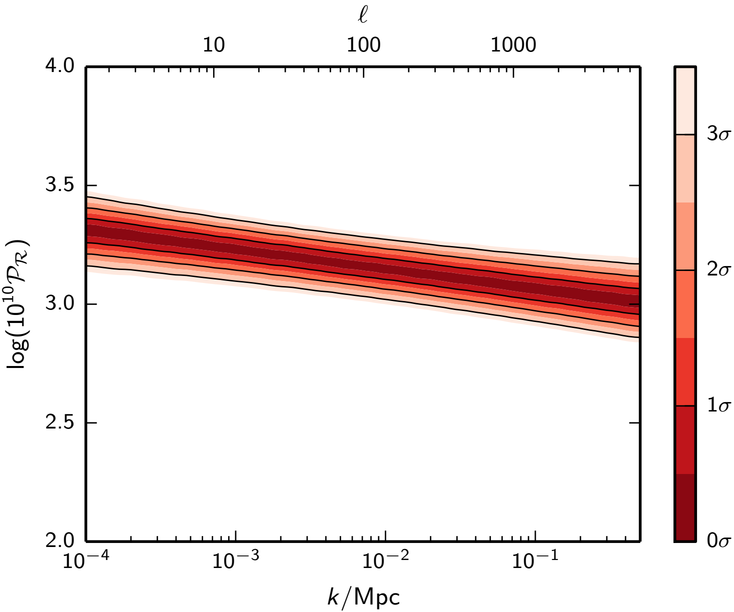
PolyChord in action

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



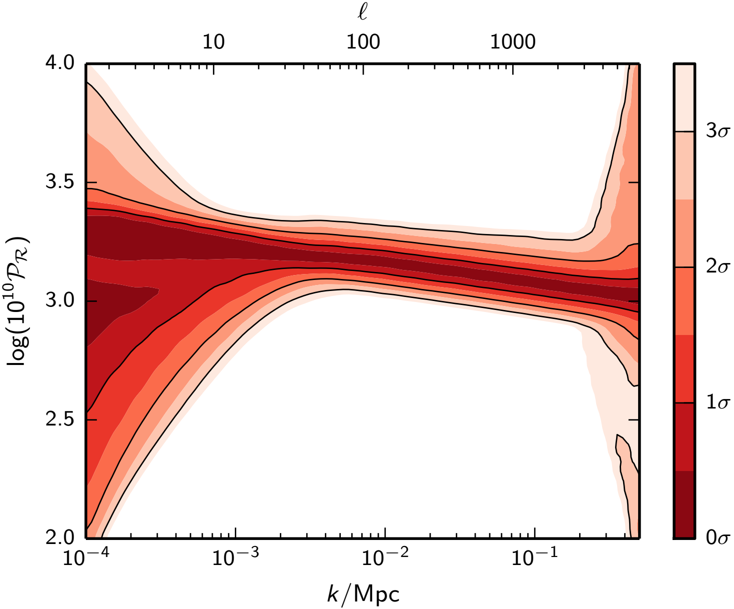
0 internal knots

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



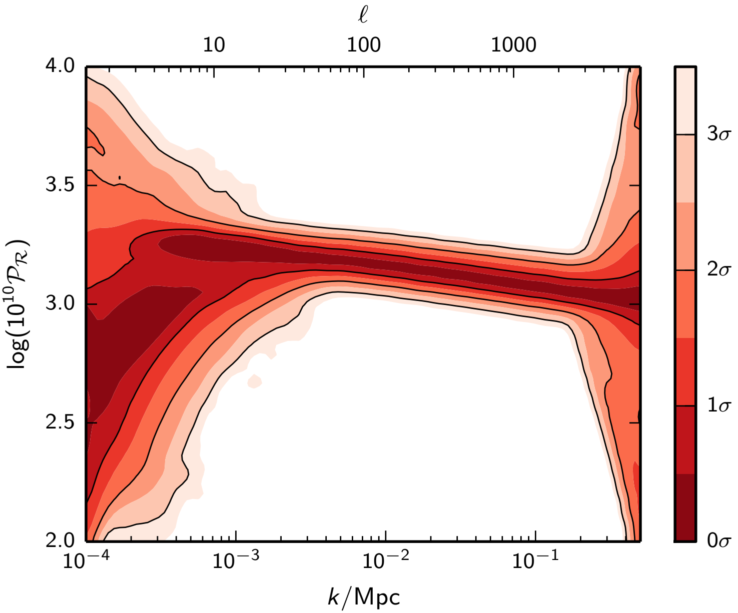
1 internal knots

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



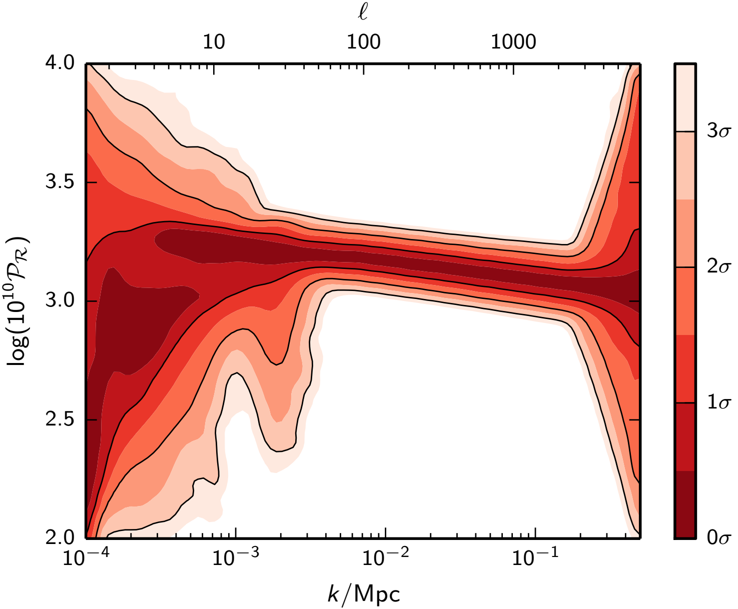
2 internal knots

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



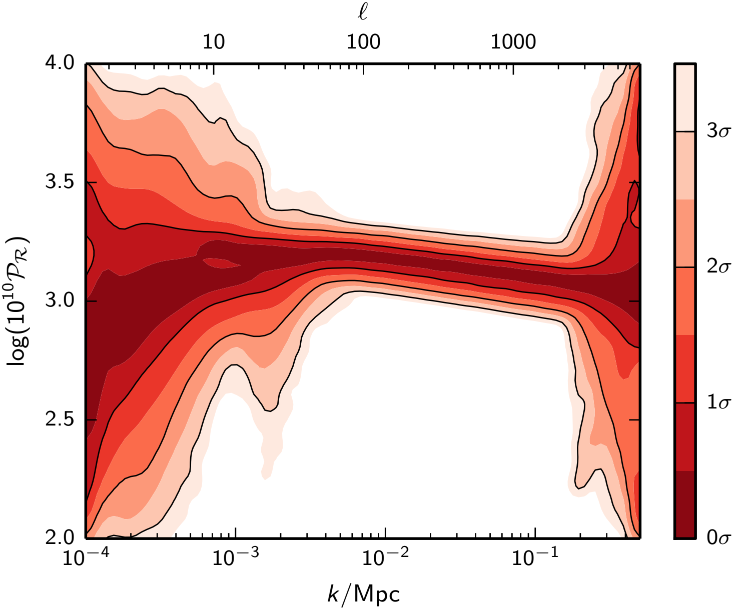
3 internal knots

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



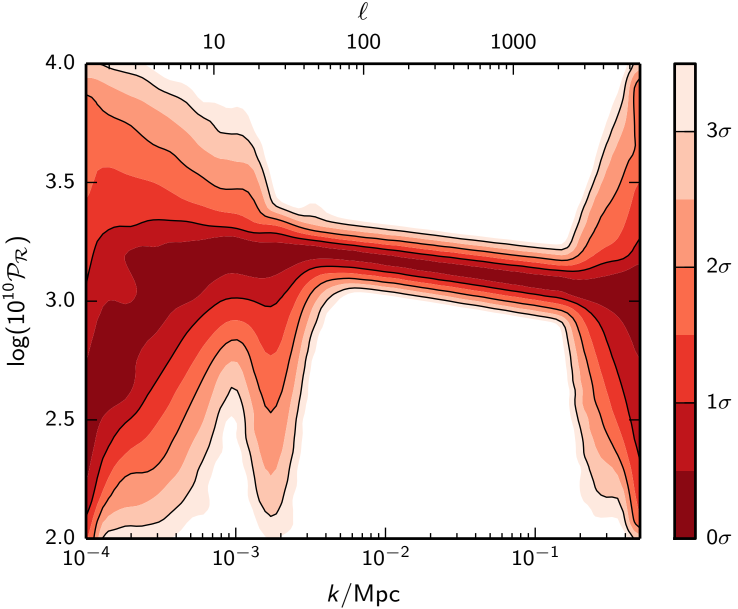
4 internal knots

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



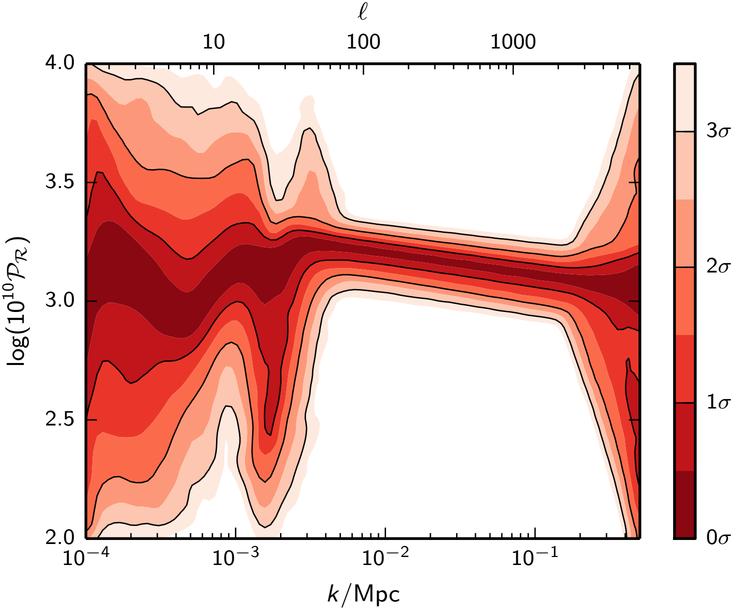
5 internal knots

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



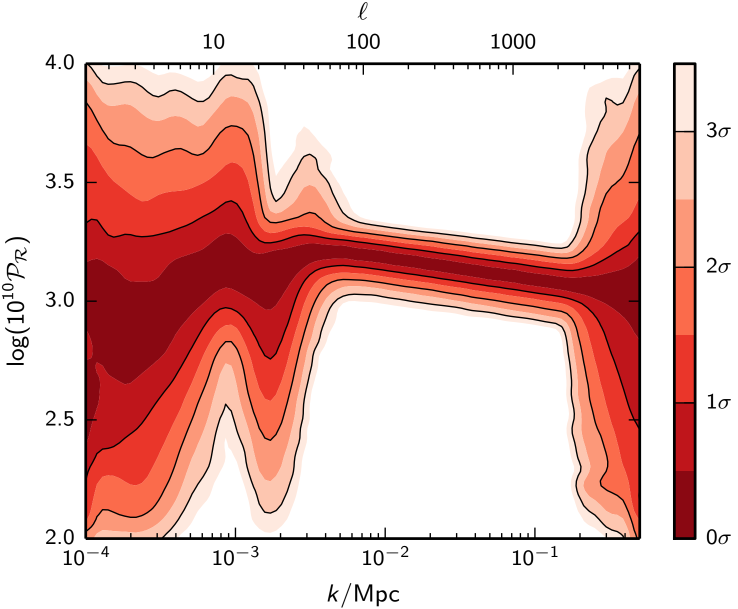
6 internal knots

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



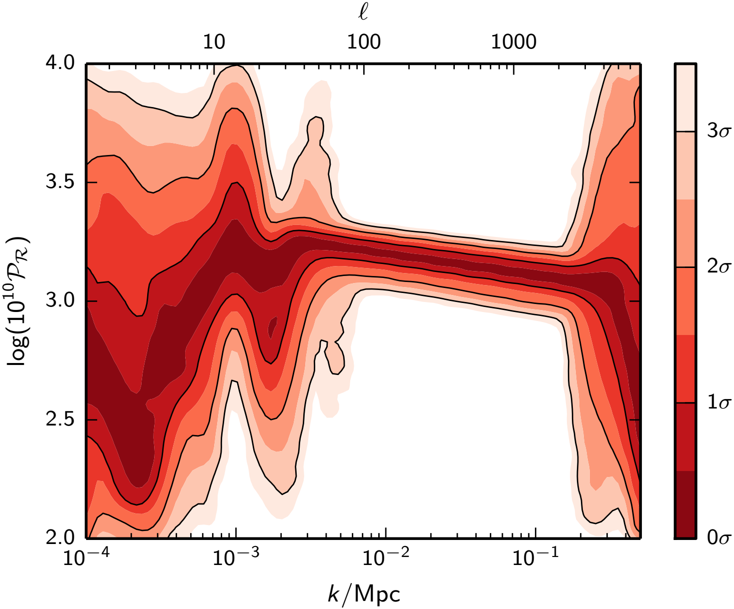
7 internal knots

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



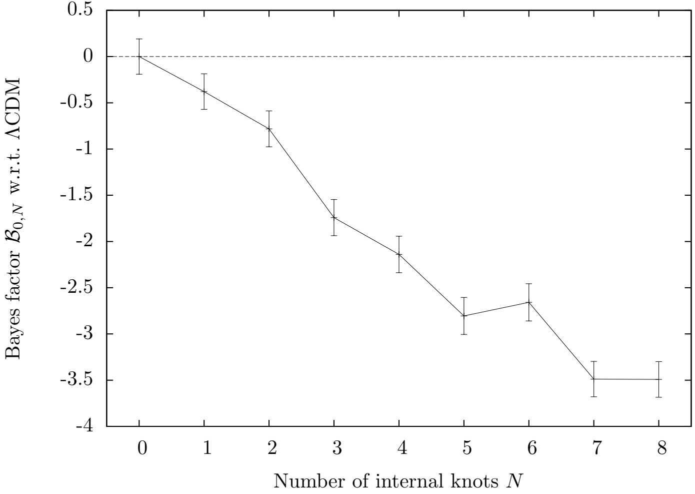
8 internal knots

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



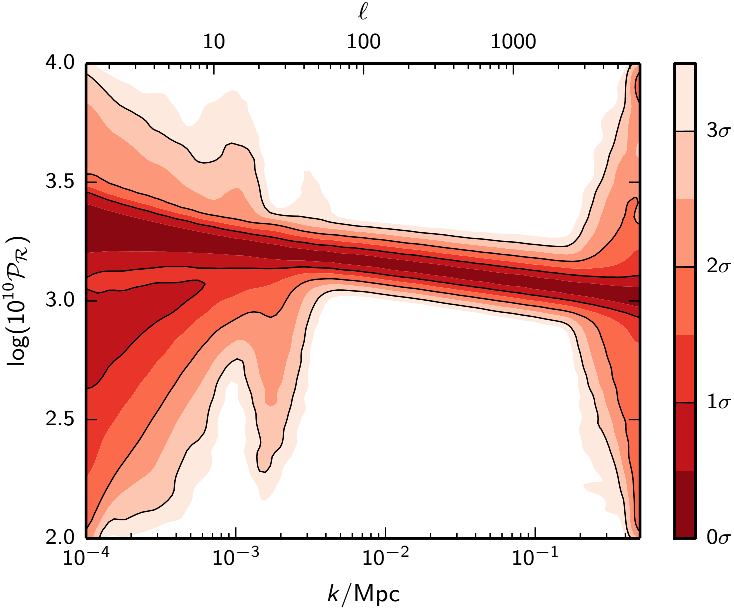
Bayes Factors

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



Marginalised plot

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



Dark energy equation of state reconstruction

Dark energy equation of state reconstruction

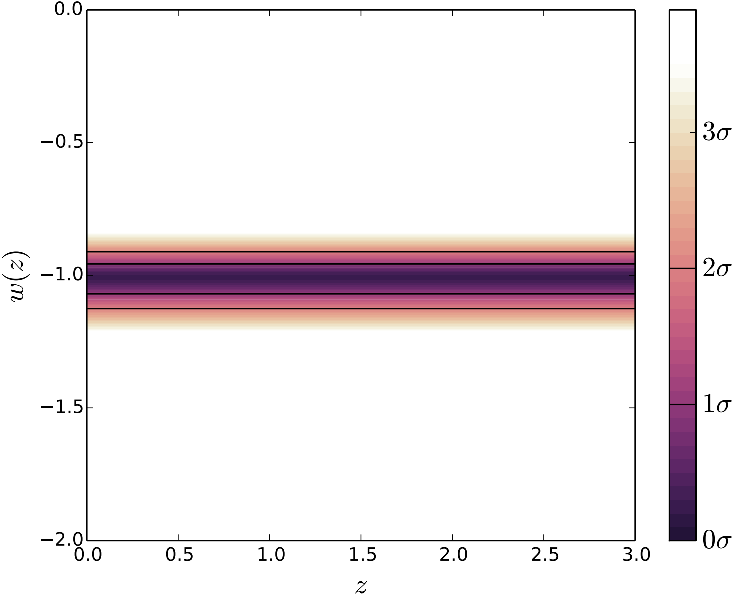
- Same thing, but for Dark energy equation of state $w(z)$ (quintessence).

Dark energy equation of state reconstruction

- Same thing, but for Dark energy equation of state $w(z)$ (quintessence).
- Data used is Planck 2015, BOSS DR 11, JLA supernovae and BOSS Lyα data

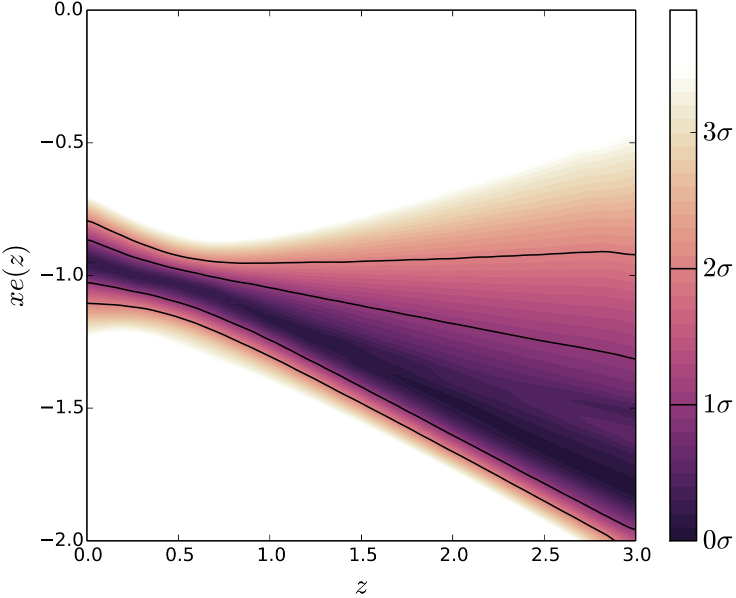
Flat, variable w

Dark energy equation of state reconstruction



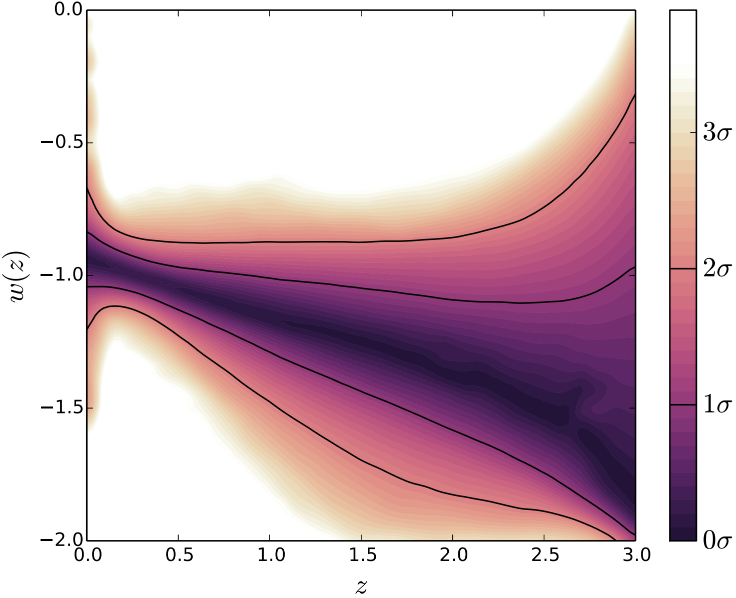
Tilted

Dark energy equation of state reconstruction



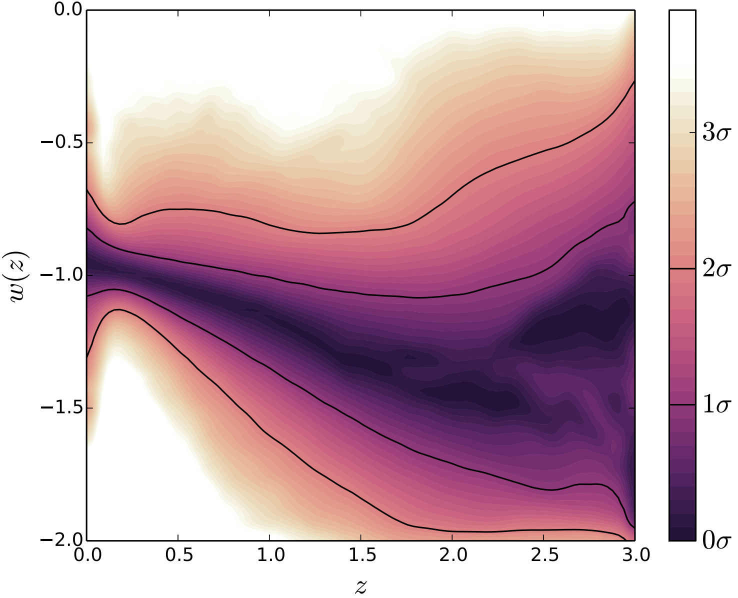
1 internal node

Dark energy equation of state reconstruction



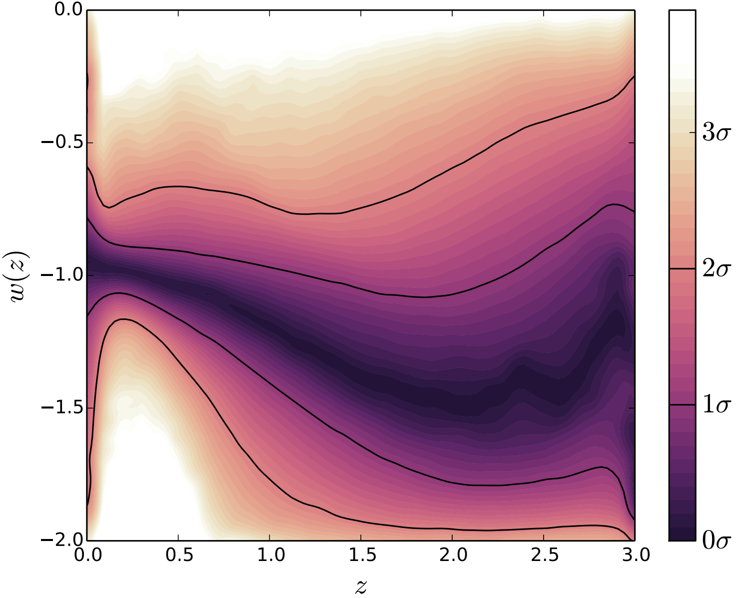
2 internal nodes

Dark energy equation of state reconstruction



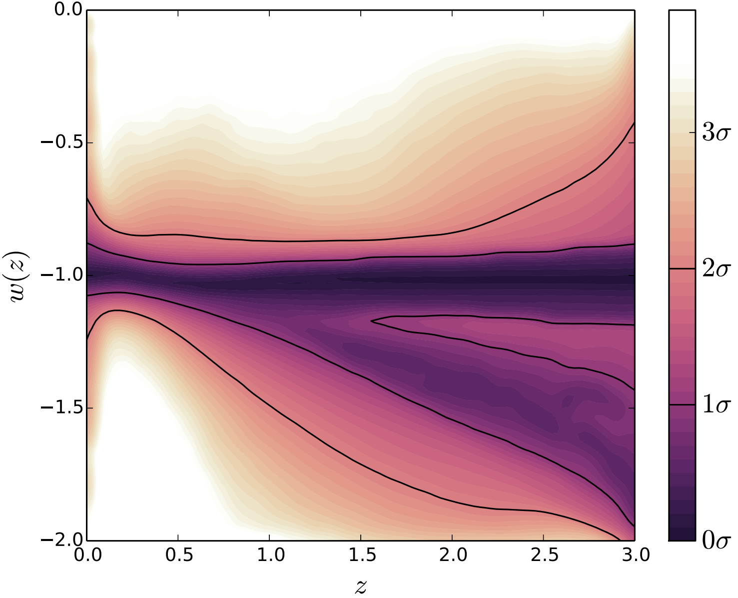
3 internal nodes

Dark energy equation of state reconstruction



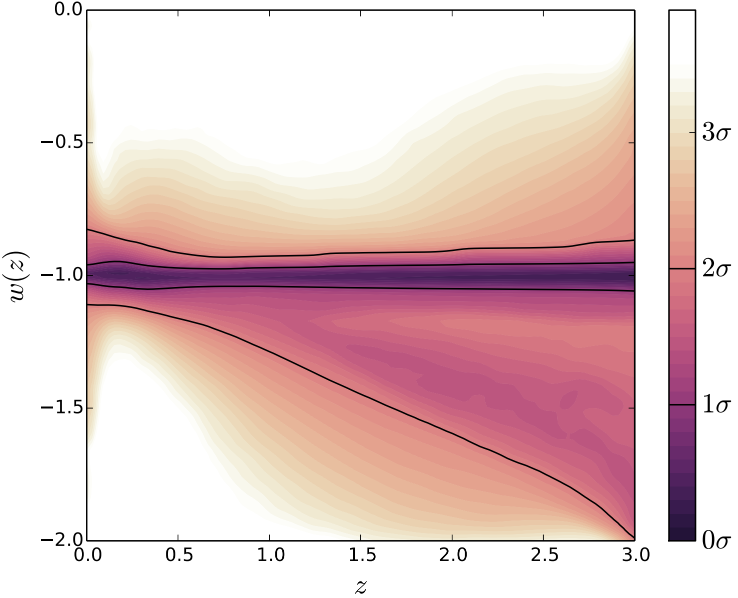
Marginalised plot - just extension models

Dark energy equation of state reconstruction



Marginalised plot - including LCDM

Dark energy equation of state reconstruction



Useful links

My email:	wh260@cam.ac.uk
My room:	Room 104, Tuesday-Thursday this week
PolyChord:	ccpforge.cse.rl.ac.uk/gf/project/polychord
MultiNest:	ccpforge.cse.rl.ac.uk/gf/project/multinest
Stan:	mc-stan.org/
emcee:	dan.iel.fm/emcee/current/