Nested Sampling

An efficient and robust Bayesian inference tool for Machine Learning and Data Science

Will Handley
wh260@cam.ac.uk

Kavli Institute for Cosmology
Astrophysics Group
Cavendish Laboratory
University of Cambridge

November 14^{th}, 2019

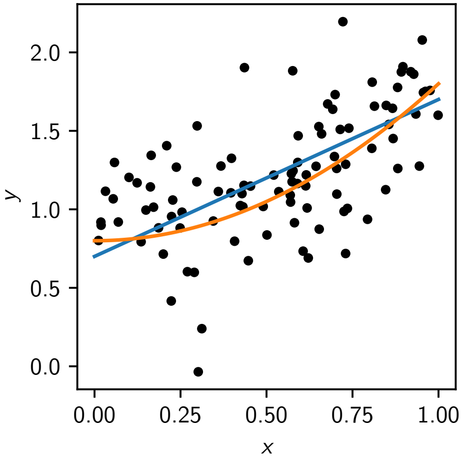
Motivating example

Fitting lines to data

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

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

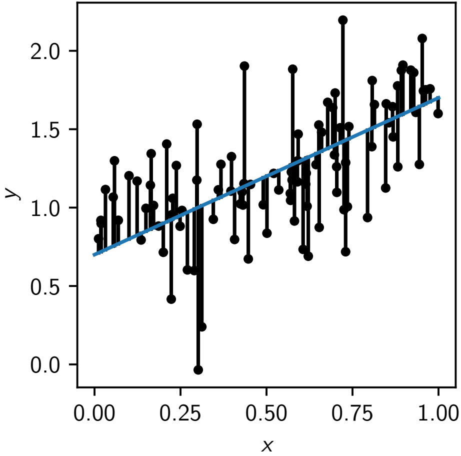
- Model parameters
  $\theta = (a, b)$



- For each parameter set θ:

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

- Minimise χ^2 wrt θ

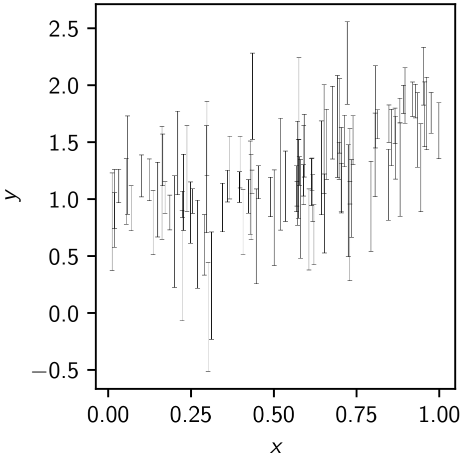


χ^2 with non-uniform data errors

Fitting lines to data

- If data have non-uniform errors:

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



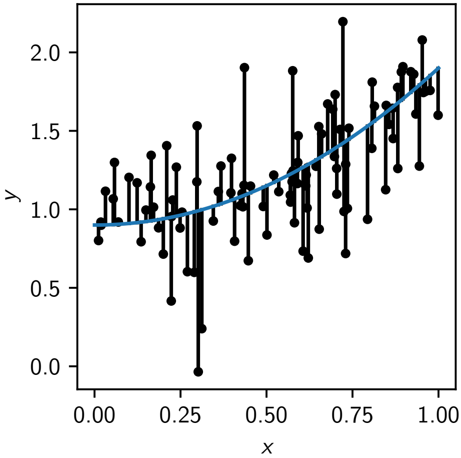
Problems with χ^2

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 approach even come from?

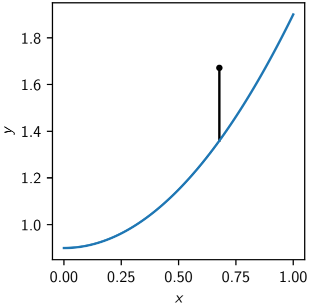


The probability of observing a datum:

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

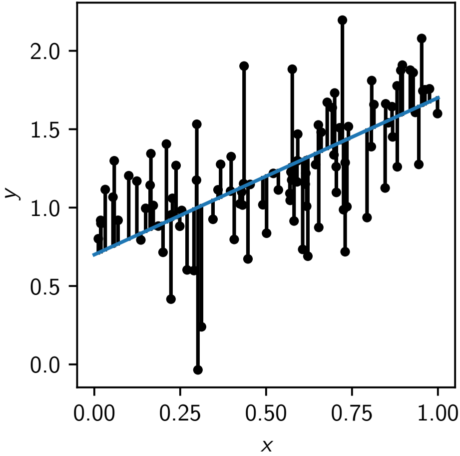
The probability of observing the data:

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



Maximum likelihood

Fitting lines to data



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

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

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

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

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

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

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:

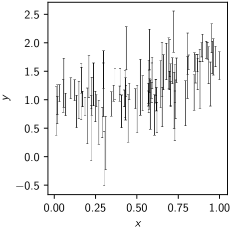
$$P(x|y)P(y) = P(x,y)$$

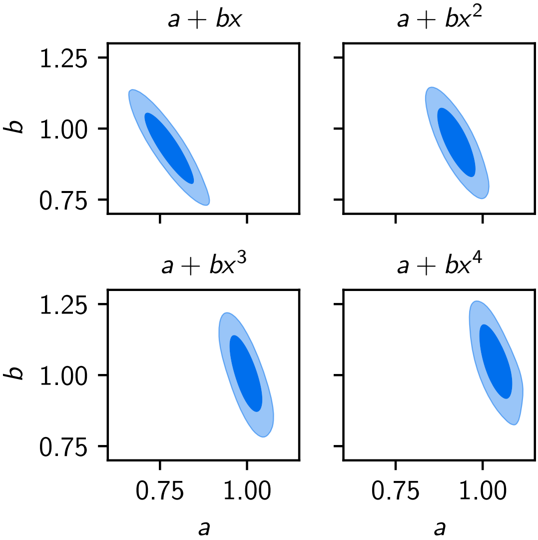
- Bayes theorem:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

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

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



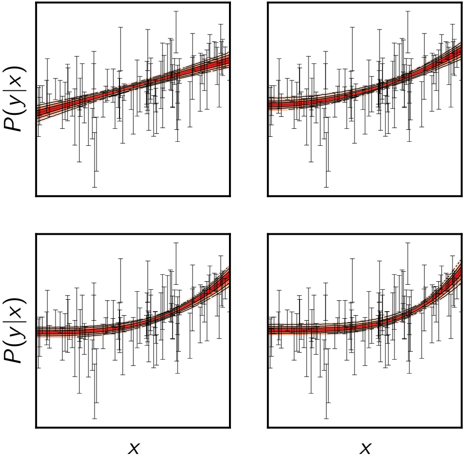


Predictive posterior

More useful to plot:

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

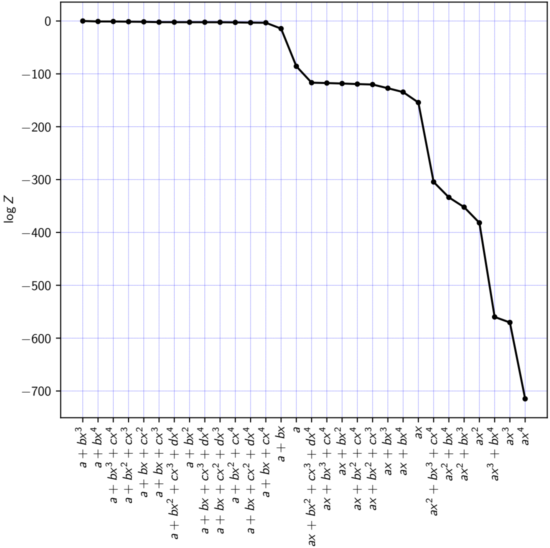
(all conditioned on D, M)



Model comparison

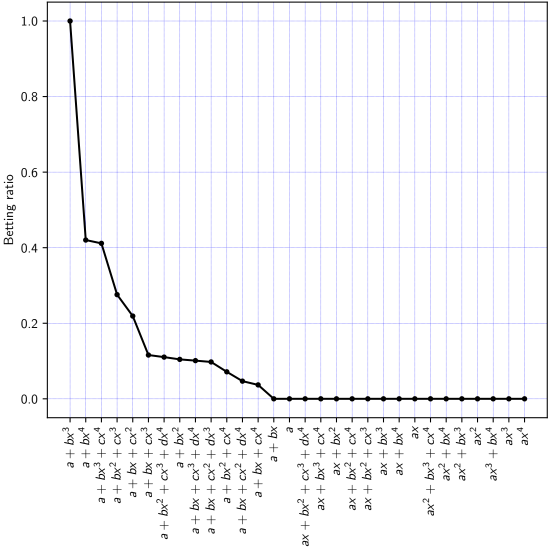
Bayesian inference

- We may use the Bayesian evidence Z to determine whether a model is reasonable.
- $Z = P(D|M) = \int P(D|M,\theta)P(\theta|M)d\theta$
- Normally assume uniform model priors $Z \propto P(M|D)P(M)$.



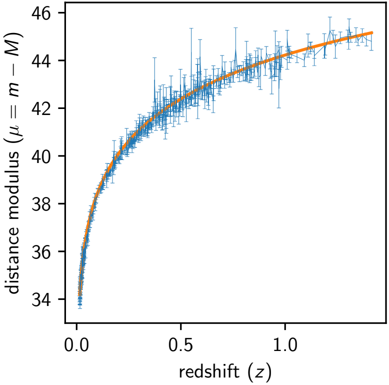
Model comparison
Bayesian inference

- We may use the Bayesian evidence Z to determine whether a model is reasonable.
- $Z = P(D|M) = \int P(D|M,\theta)P(\theta|M)d\theta$
- Normally assume uniform model priors $Z \propto P(M|D)P(M)$.



Line fitting (context)

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



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

Quantifying error with Probability

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

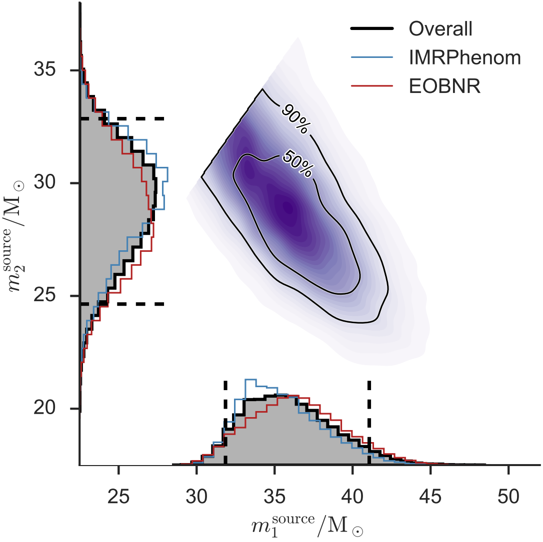
$$13.73 \pm 0.12\,\text{billion years old.}$$

- 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. 66% confident of the value lying in this range.
- More importantly, these are *summary statistics*.

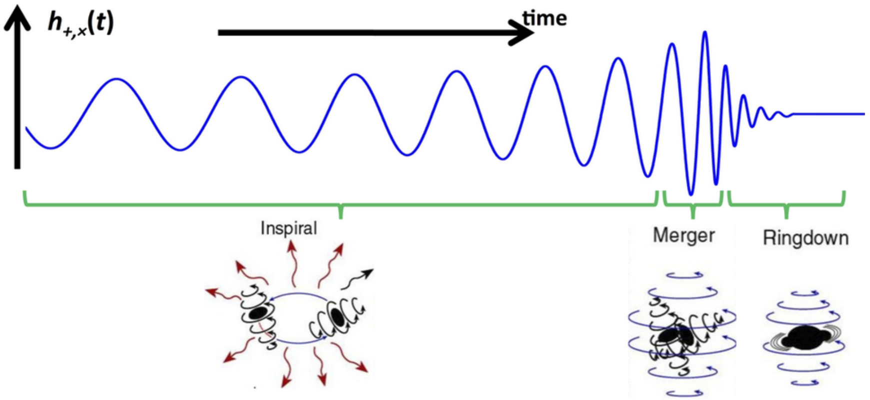
LIGO binary merger



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

Theory

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)

Posterior $\mathcal{P}$

Extended example of inference: LIGO

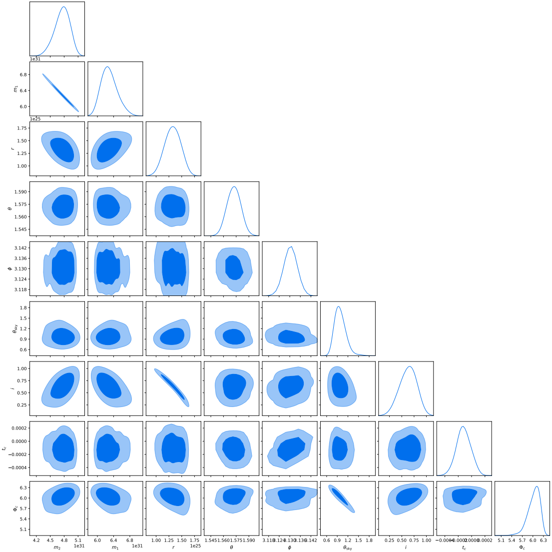
- Cannot plot the full posterior distribution:

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

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

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

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

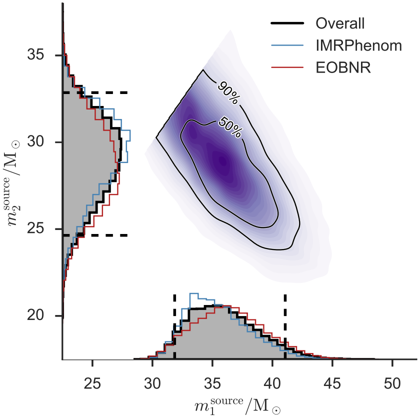


- Does give insight
- Not the full picture

Sampling

How to describe a high-dimensional posterior

- In high dimensions, posterior $\mathcal{P}$ occupies a vanishingly small region of the prior π.
- Gridding is doomed to failure for $D \gtrsim 4$.
- *Sampling* the posterior is an excellent compression scheme.



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:
$$\bar{m}_1 \equiv \langle m_1 \rangle_{\mathcal{P}} \equiv \int m_1 P(\theta|D, M) d\theta$$
- Mass covariance:
$$\mathrm{Cov}(m_1, m_2) \equiv \int (m_1 - \bar{m}_1)(m_2 - \bar{m}_2) P(\theta|D, M) d\theta$$
- 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:
$$\mathrm{Cov}(m_1, m_2) \approx \frac{\sum_{i=1}^{N} w^{(i)} (m_1^{(i)} - \bar{m}_1)(m_2^{(i)} - \bar{m}_2)}{\sum_{i=1}^{N} w^{(i)}}$$
- Marginalised samples: Just ignore the other coordinates.
- N.B. Typically have *weighted* samples

Cosmology in high dimensions

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

Cosmology in high dimensions

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

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

Cosmology in high dimensions

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

$$D = \{C_\ell\}$$
$$M = \Lambda\text{CDM}$$

Cosmology in high dimensions

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

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

Cosmology in high dimensions

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

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

Cosmology in high dimensions

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

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

Cosmology in high dimensions

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

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

Cosmology in high dimensions

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

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

Cosmology in high dimensions

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

$$\begin{aligned}
D =& \{C_\ell^{(\mathrm{Planck})}\} \\
M =& \Lambda\mathrm{CDM} + \text{extensions} \\
\Theta =& \Theta_{\Lambda\mathrm{CDM}} + \Theta_{\mathrm{Planck}} \\
\Theta_{\Lambda\mathrm{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10}A_s), n_s) \\
\Theta_{\mathrm{Planck}} =& (y_{\mathrm{cal}}, A_{217}^{CIB}, \xi^{tSZ-CIB}, A_{143}^{tSZ}, A_{100}^{PS}, A_{143}^{PS}, A_{143\times217}^{PS}, A_{217}^{PS}, A^{kSZ}, \\
& A_{100}^{\mathrm{dust}\,TT}, A_{143}^{\mathrm{dust}\,TT}, A_{143\times217}^{\mathrm{dust}\,TT}, A_{217}^{\mathrm{dust}\,TT}, A_{100}^{\mathrm{dust}\,TE}, A_{100\times143}^{\mathrm{dust}\,TE}, \\
& A_{100\times217}^{\mathrm{dust}\,TE}, A_{143}^{\mathrm{dust}\,TE}, A_{143\times217}^{\mathrm{dust}\,TE}, A_{217}^{\mathrm{dust}\,TE}, c_{100}, c_{217})
\end{aligned}$$

Cosmology in high dimensions

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

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

Cosmology in high dimensions

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

$$\begin{aligned}
D =& \{C_\ell^{(\text{Planck})}\} \\
M =& \Lambda\text{CDM} + \text{extensions} \\
\Theta =& \Theta_{\Lambda\text{CDM}} + \Theta_{\text{Planck}} + \Theta_{\text{extensions}} \\
\Theta_{\Lambda\text{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10} A_s), n_s) \\
\Theta_{\text{Planck}} =& (y_{\mathrm{cal}}, A^{CIB}_{217}, \xi^{tSZ-CIB}, A^{tSZ}_{143}, A^{PS}_{100}, A^{PS}_{143}, A^{PS}_{143\times217}, A^{PS}_{217}, A^{kSZ}, \\
& A^{\mathrm{dust}\,TT}_{100}, A^{\mathrm{dust}\,TT}_{143}, A^{\mathrm{dust}\,TT}_{143\times217}, A^{\mathrm{dust}\,TT}_{217}, A^{\mathrm{dust}\,TE}_{100}, A^{\mathrm{dust}\,TE}_{100\times143}, \\
& A^{\mathrm{dust}\,TE}_{100\times217}, A^{\mathrm{dust}\,TE}_{143}, A^{\mathrm{dust}\,TE}_{143\times217}, A^{\mathrm{dust}\,TE}_{217}, c_{100}, c_{217}) \\
\Theta_{\text{extensions}} =& (n_{\mathrm{run}})
\end{aligned}$$

Cosmology in high dimensions

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

$$\begin{aligned}
D =& \{C_\ell^{(\mathrm{Planck})}\} \\
M =& \Lambda\mathrm{CDM} + \text{extensions} \\
\Theta =& \Theta_{\Lambda\mathrm{CDM}} + \Theta_{\mathrm{Planck}} + \Theta_{\mathrm{extensions}} \\
\Theta_{\Lambda\mathrm{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10} A_s), n_s) \\
\Theta_{\mathrm{Planck}} =& (y_{\mathrm{cal}}, A^{CIB}_{217}, \xi^{tSZ-CIB}, A^{tSZ}_{143}, A^{PS}_{100}, A^{PS}_{143}, A^{PS}_{143\times217}, A^{PS}_{217}, A^{kSZ}, \\
& A^{\mathrm{dust}\,TT}_{100}, A^{\mathrm{dust}\,TT}_{143}, A^{\mathrm{dust}\,TT}_{143\times217}, A^{\mathrm{dust}\,TT}_{217}, A^{\mathrm{dust}\,TE}_{100}, A^{\mathrm{dust}\,TE}_{100\times143}, \\
& A^{\mathrm{dust}\,TE}_{100\times217}, A^{\mathrm{dust}\,TE}_{143}, A^{\mathrm{dust}\,TE}_{143\times217}, A^{\mathrm{dust}\,TE}_{217}, c_{100}, c_{217}) \\
\Theta_{\mathrm{extensions}} =& (n_{\mathrm{run}}, n_{\mathrm{run,run}})
\end{aligned}$$

Cosmology in high dimensions

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

$$\begin{aligned}
D =& \{C_\ell^{(\mathrm{Planck})}\} \\
M =& \Lambda\mathrm{CDM} + \text{extensions} \\
\Theta =& \Theta_{\Lambda\mathrm{CDM}} + \Theta_{\mathrm{Planck}} + \Theta_{\mathrm{extensions}} \\
\Theta_{\Lambda\mathrm{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10}A_s), n_s) \\
\Theta_{\mathrm{Planck}} =& (y_{\mathrm{cal}}, A_{217}^{CIB}, \xi^{tSZ-CIB}, A_{143}^{tSZ}, A_{100}^{PS}, A_{143}^{PS}, A_{143\times217}^{PS}, A_{217}^{PS}, A^{kSZ}, \\
& A_{100}^{\mathrm{dust}\,TT}, A_{143}^{\mathrm{dust}\,TT}, A_{143\times217}^{\mathrm{dust}\,TT}, A_{217}^{\mathrm{dust}\,TT}, A_{100}^{\mathrm{dust}\,TE}, A_{100\times143}^{\mathrm{dust}\,TE}, \\
& A_{100\times217}^{\mathrm{dust}\,TE}, A_{143}^{\mathrm{dust}\,TE}, A_{143\times217}^{\mathrm{dust}\,TE}, A_{217}^{\mathrm{dust}\,TE}, c_{100}, c_{217}) \\
\Theta_{\mathrm{extensions}} =& (n_{\mathrm{run}}, n_{\mathrm{run,run}}, w)
\end{aligned}$$

Cosmology in high dimensions

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

$$\begin{aligned}
D =& \{C_\ell^{(\text{Planck})}\} \\
M =& \Lambda\text{CDM} + \text{extensions} \\
\Theta =& \Theta_{\Lambda\text{CDM}} + \Theta_{\text{Planck}} + \Theta_{\text{extensions}} \\
\Theta_{\Lambda\text{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10} A_s), n_s) \\
\Theta_{\text{Planck}} =& (y_{\mathrm{cal}}, A_{217}^{CIB}, \xi^{tSZ-CIB}, A_{143}^{tSZ}, A_{100}^{PS}, A_{143}^{PS}, A_{143\times217}^{PS}, A_{217}^{PS}, A^{kSZ}, \\
& A_{100}^{\mathrm{dust}\,TT}, A_{143}^{\mathrm{dust}\,TT}, A_{143\times217}^{\mathrm{dust}\,TT}, A_{217}^{\mathrm{dust}\,TT}, A_{100}^{\mathrm{dust}\,TE}, A_{100\times143}^{\mathrm{dust}\,TE}, \\
& A_{100\times217}^{\mathrm{dust}\,TE}, A_{143}^{\mathrm{dust}\,TE}, A_{143\times217}^{\mathrm{dust}\,TE}, A_{217}^{\mathrm{dust}\,TE}, c_{100}, c_{217}) \\
\Theta_{\text{extensions}} =& (n_{\mathrm{run}}, n_{\mathrm{run,run}}, w, \Sigma m_\nu, m_{\nu,\mathrm{sterile}}^{\mathrm{eff}})
\end{aligned}$$

Cosmology in high dimensions

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

$$\begin{aligned}
D =& \{C_\ell^{(\text{Planck})}\} + \{\text{LSS}\} \\
M =& \Lambda\text{CDM} + \text{extensions} \\
\Theta =& \Theta_{\Lambda\text{CDM}} + \Theta_{\text{Planck}} + \Theta_{\text{extensions}} \\
\Theta_{\Lambda\text{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10}A_s), n_s) \\
\Theta_{\text{Planck}} =& (y_{\text{cal}}, A_{217}^{CIB}, \xi^{tSZ-CIB}, A_{143}^{tSZ}, A_{100}^{PS}, A_{143}^{PS}, A_{143\times217}^{PS}, A_{217}^{PS}, A^{kSZ}, \\
& A_{100}^{\text{dust}\,TT}, A_{143}^{\text{dust}\,TT}, A_{143\times217}^{\text{dust}\,TT}, A_{217}^{\text{dust}\,TT}, A_{100}^{\text{dust}\,TE}, A_{100\times143}^{\text{dust}\,TE}, \\
& A_{100\times217}^{\text{dust}\,TE}, A_{143}^{\text{dust}\,TE}, A_{143\times217}^{\text{dust}\,TE}, A_{217}^{\text{dust}\,TE}, c_{100}, c_{217}) \\
\Theta_{\text{extensions}} =& (n_{\text{run}}, n_{\text{run,run}}, w, \Sigma m_\nu, m_{\nu,\text{sterile}}^{\text{eff}})
\end{aligned}$$

Cosmology in high dimensions

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

$$\begin{aligned}
D =& \{C_\ell^{(\mathrm{Planck})}\} + \{\mathrm{LSS}\} + \{\text{“Big Data”}\} \\
M =& \Lambda\mathrm{CDM} + \text{extensions} \\
\Theta =& \Theta_{\Lambda\mathrm{CDM}} + \Theta_{\mathrm{Planck}} + \Theta_{\mathrm{extensions}} \\
\Theta_{\Lambda\mathrm{CDM}} =& (\Omega_b h^2, \Omega_c h^2, 100\theta_{MC}, \tau, \ln(10^{10}A_s), n_s) \\
\Theta_{\mathrm{Planck}} =& (y_{\mathrm{cal}}, A^{CIB}_{217}, \xi^{tSZ-CIB}, A^{tSZ}_{143}, A^{PS}_{100}, A^{PS}_{143}, A^{PS}_{143\times217}, A^{PS}_{217}, A^{kSZ}, \\
& A^{\mathrm{dust}\,TT}_{100}, A^{\mathrm{dust}\,TT}_{143}, A^{\mathrm{dust}\,TT}_{143\times217}, A^{\mathrm{dust}\,TT}_{217}, A^{\mathrm{dust}\,TE}_{100}, A^{\mathrm{dust}\,TE}_{100\times143}, \\
& A^{\mathrm{dust}\,TE}_{100\times217}, A^{\mathrm{dust}\,TE}_{143}, A^{\mathrm{dust}\,TE}_{143\times217}, A^{\mathrm{dust}\,TE}_{217}, c_{100}, c_{217}) \\
\Theta_{\mathrm{extensions}} =& (n_{\mathrm{run}}, n_{\mathrm{run,run}}, w, \Sigma m_\nu, m^{\mathrm{eff}}_{\nu,\mathrm{sterile}})
\end{aligned}$$

Cosmology in high dimensions

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

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

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

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.

Metropolis Hastings

- Turn the N-dimensional problem into a one-dimensional one.
  1. Propose random step
  2. If uphill, make step…
  3. …otherwise sometimes make step.
- `chi-feng.github.io/mcmc-demo/`

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.
- Thermodynamic annealing
  - Suffers from same tuning issues as MCMC
- Nearest neighbor volume estimation (Heavens arXiv:1704.03472)
  - Does not scale to high dimensions $D \gtrsim 10$.

Nested Sampling

John Skilling's alternative to traditional MCMC!

- Nested sampling is a completely different way of sampling.
- Uses ensemble sampling to compress prior to posterior.

  New procedure:

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

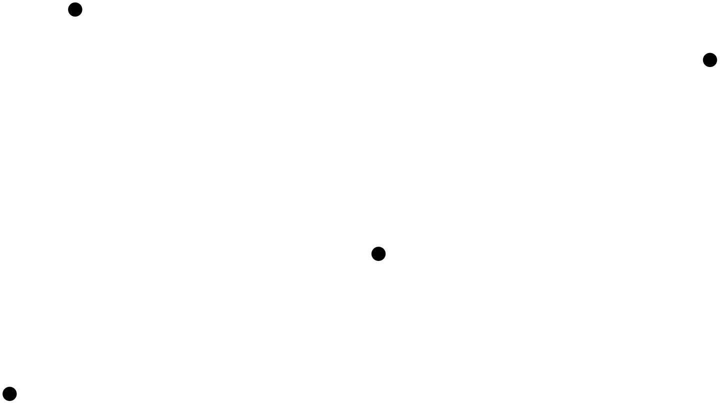
  S_0: Generate n samples uniformly over the space (from the prior π).

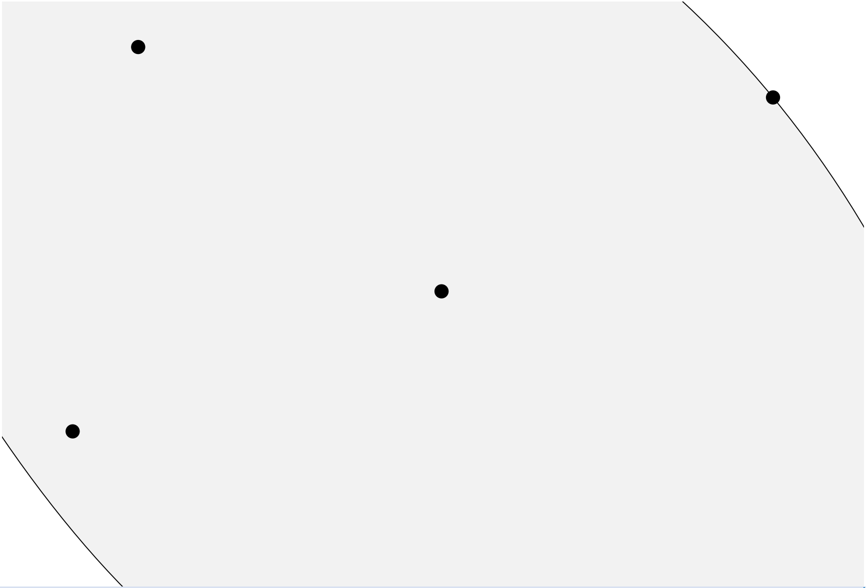
  S_{n+1}: Delete the lowest likelihood sample in S_n, and replace it with a new uniform sample with higher likelihood

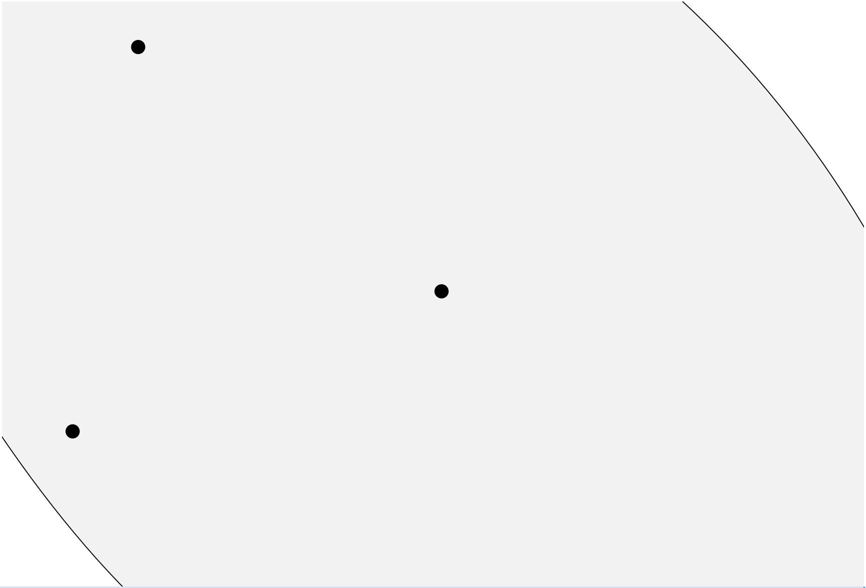
  Requires one to be able to uniformly within a region, subject to a *hard likelihood constraint*.

Nested Sampling

Graphical aid

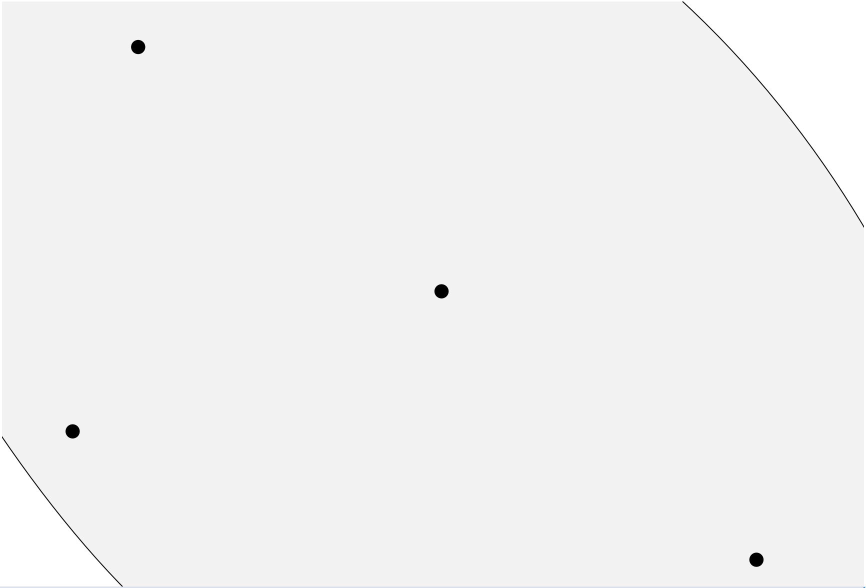






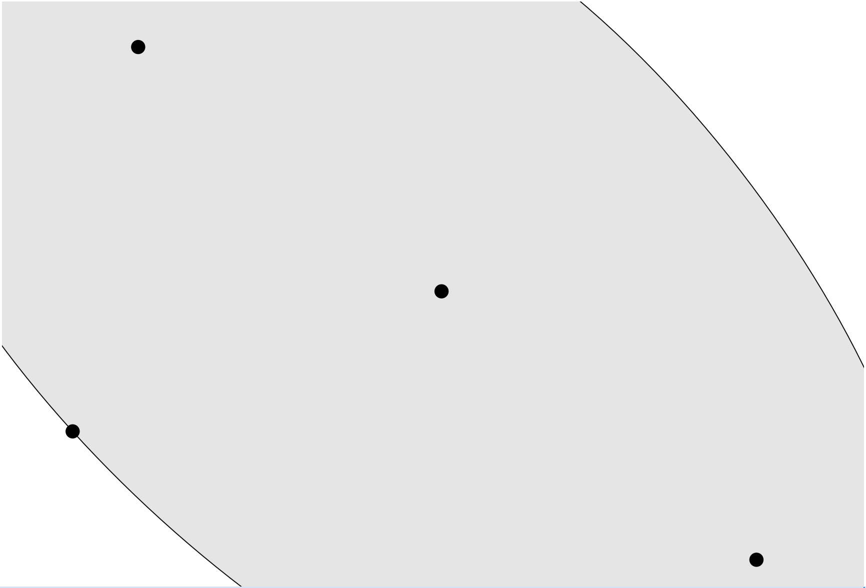
Nested Sampling

Graphical aid



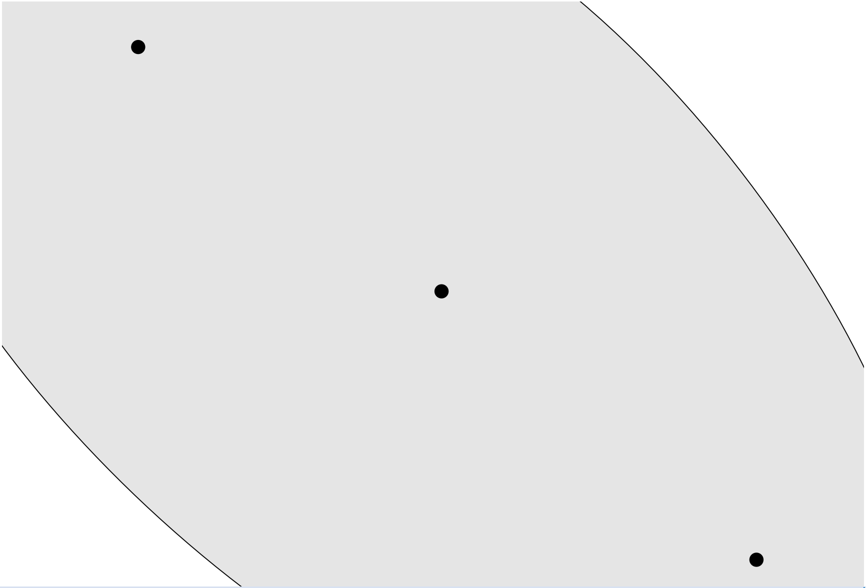
Nested Sampling

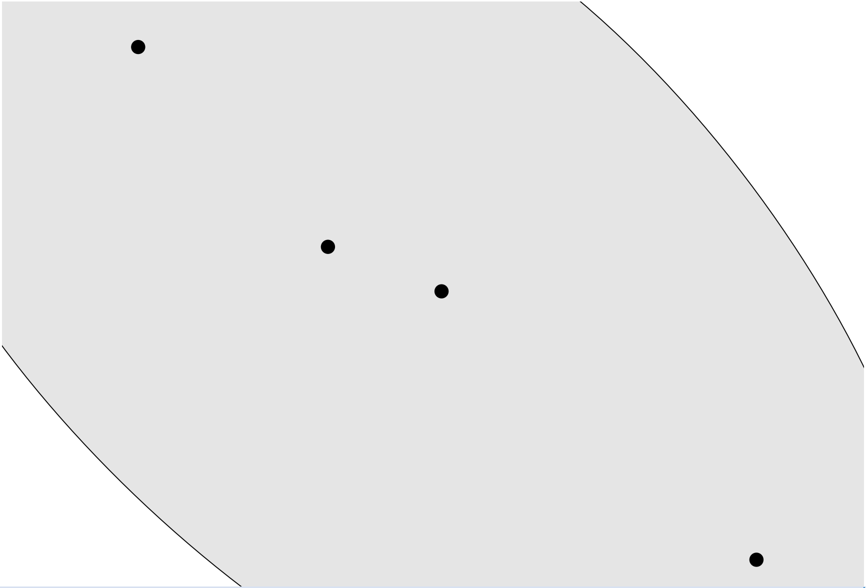
Graphical aid



Nested Sampling

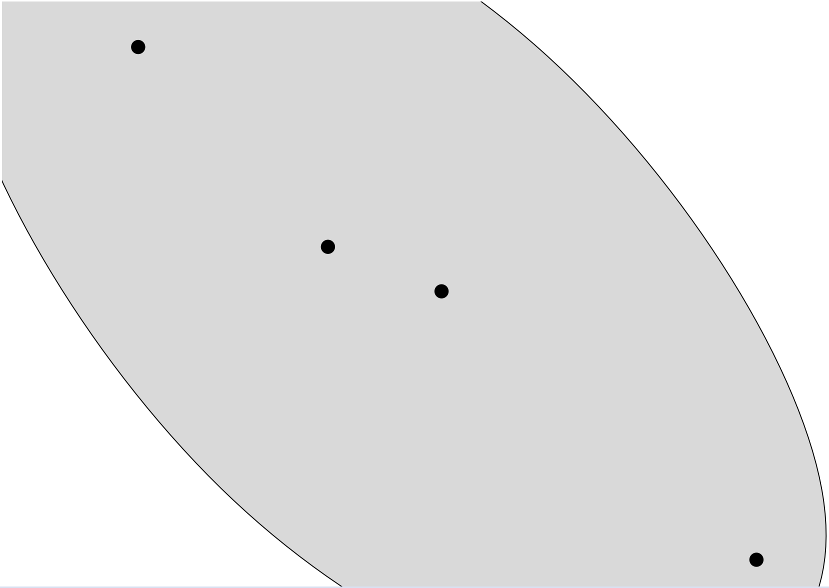
Graphical aid

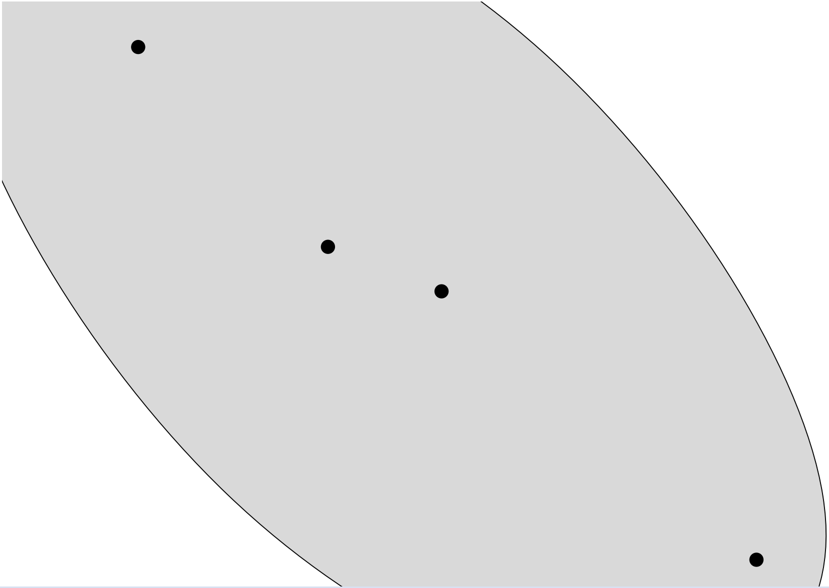




Nested Sampling

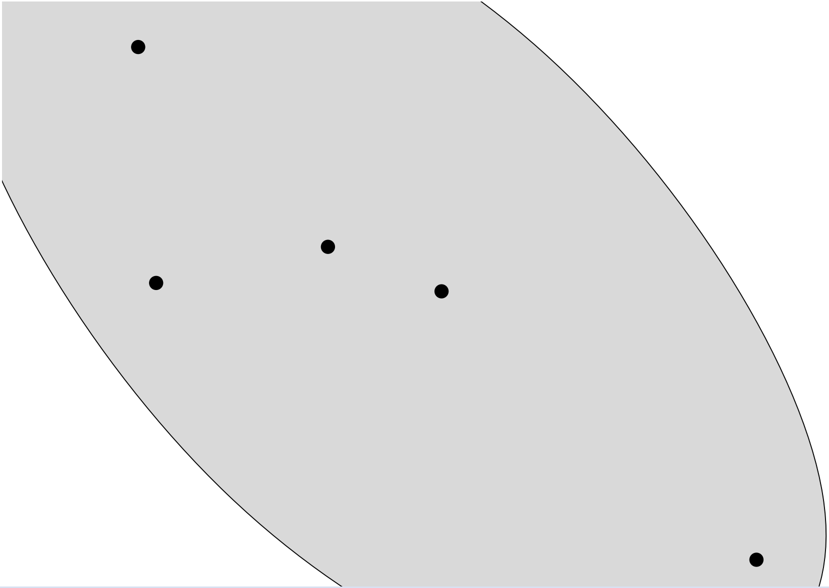
Graphical aid

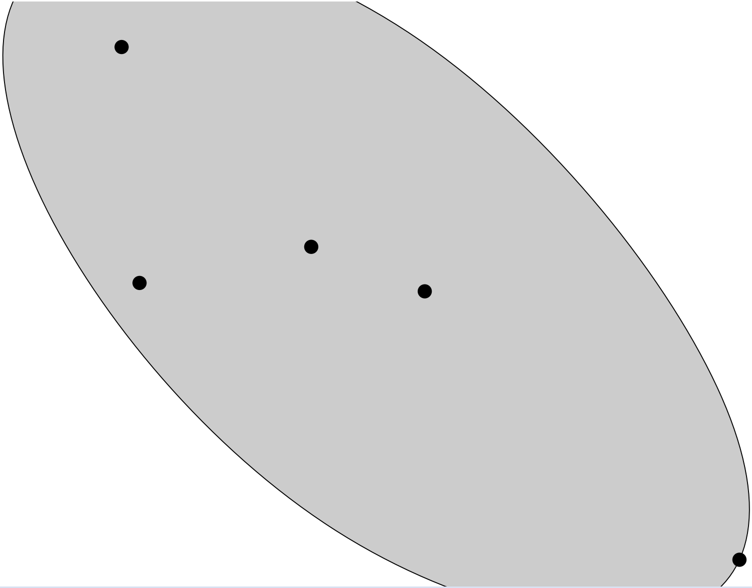


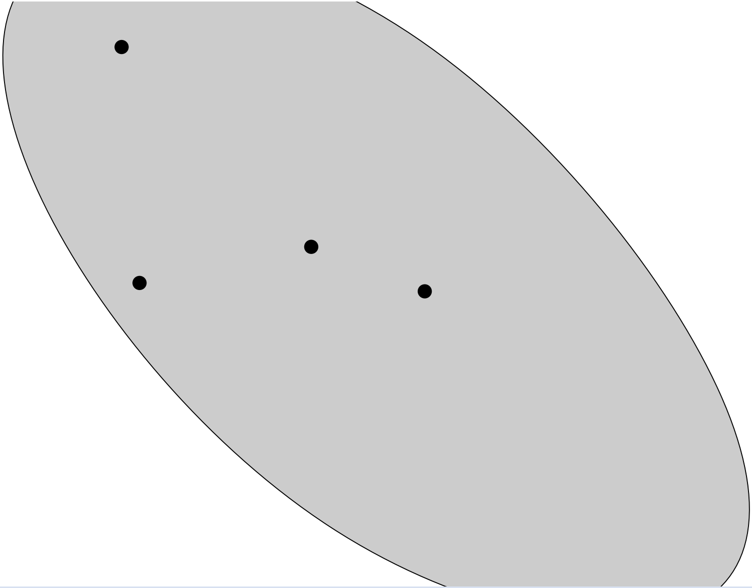


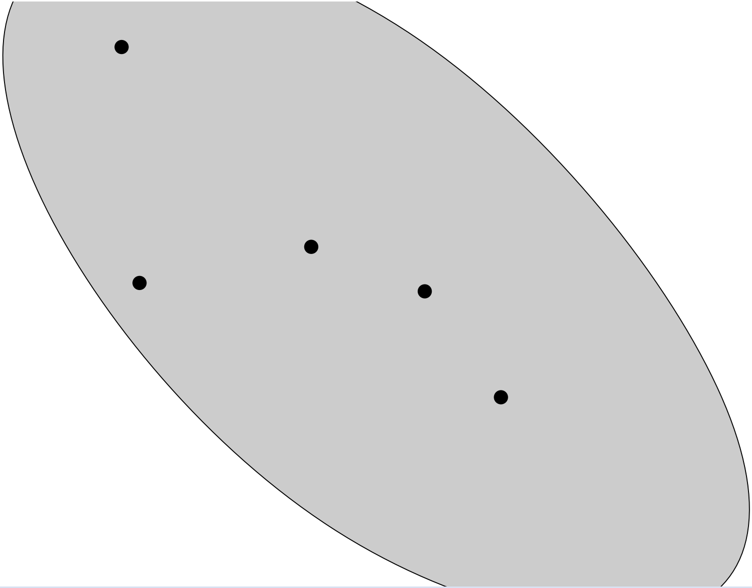
Nested Sampling

Graphical aid



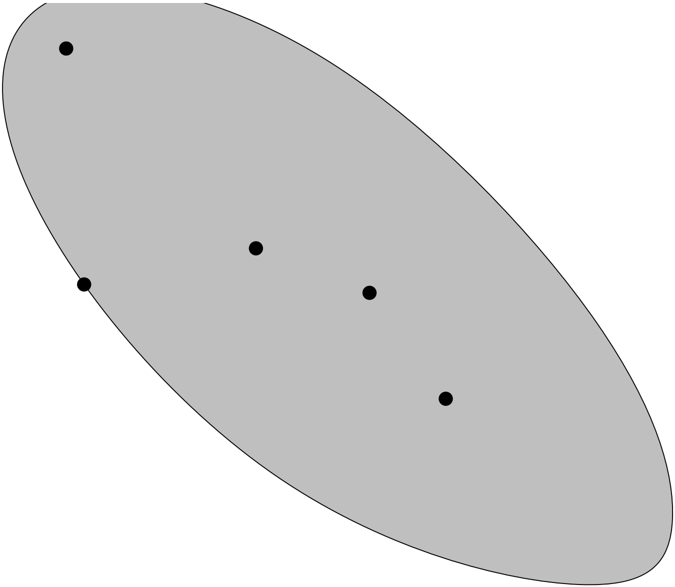






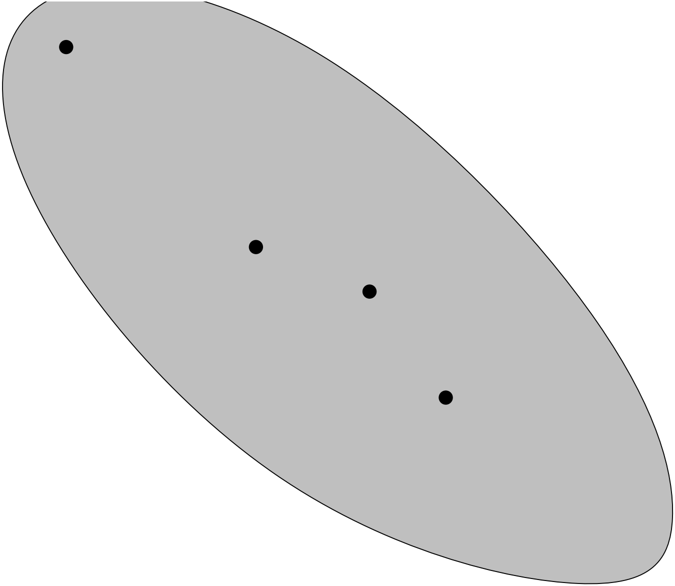
Nested Sampling

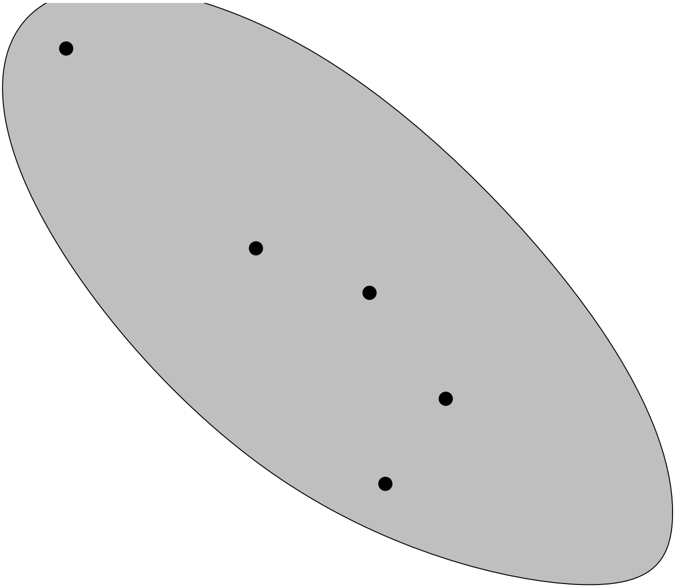
Graphical aid

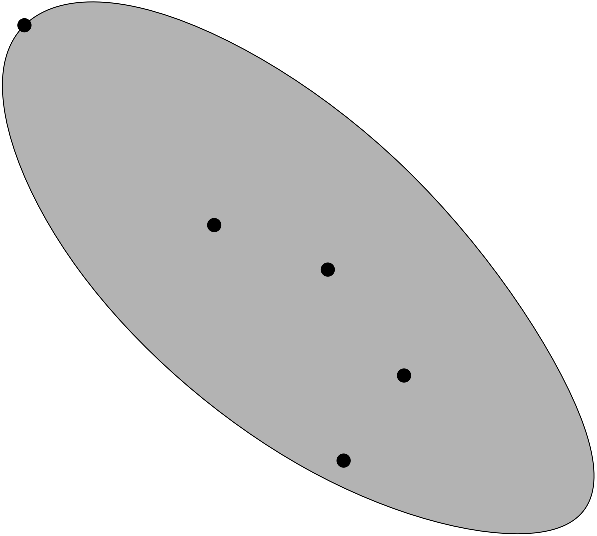


Nested Sampling

Graphical aid

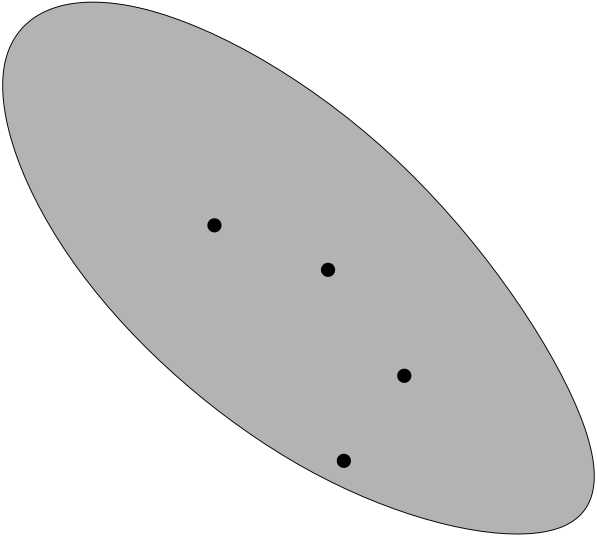






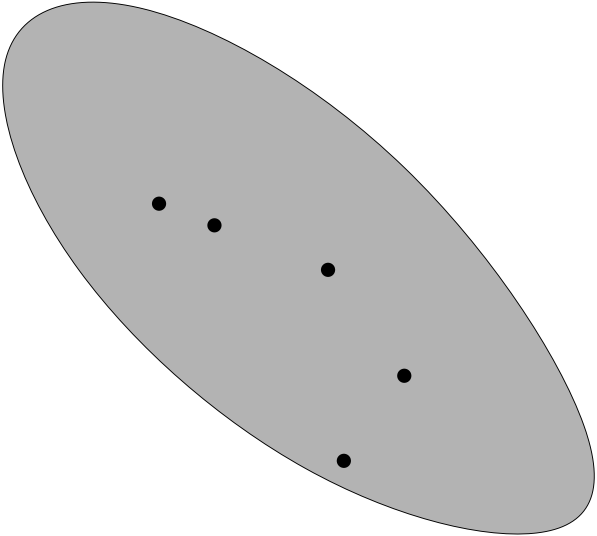
Nested Sampling

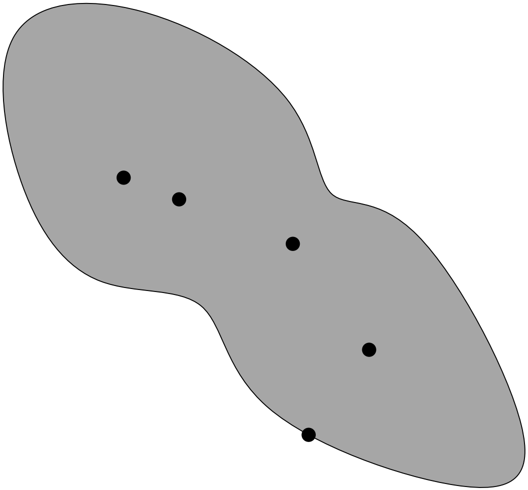
Graphical aid



Nested Sampling

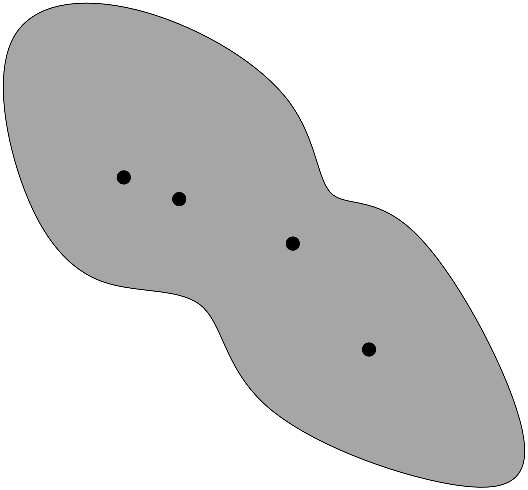
Graphical aid

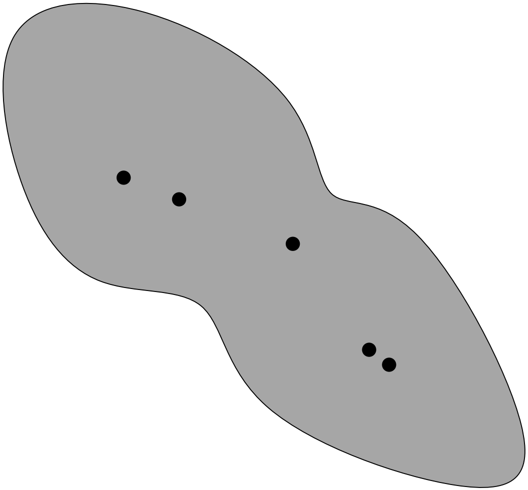




Nested Sampling

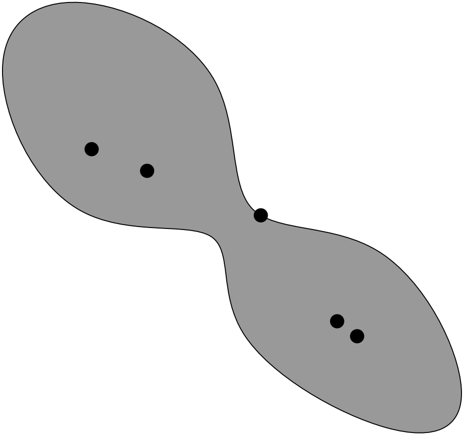
Graphical aid

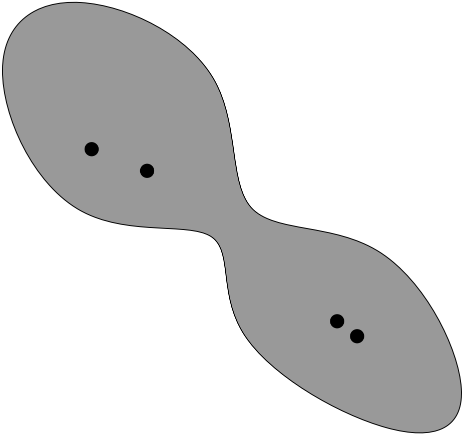




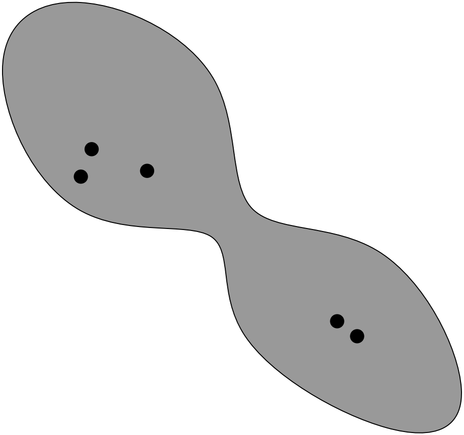
Nested Sampling

Graphical aid



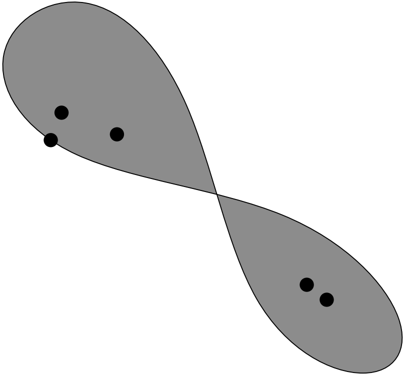


Nested Sampling
Graphical aid



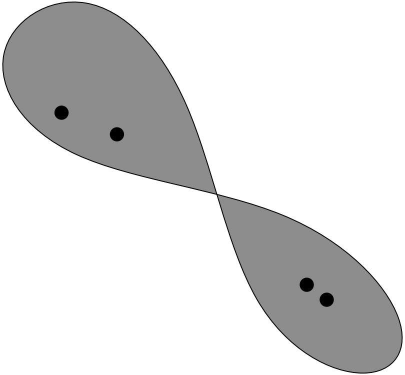
Nested Sampling

Graphical aid



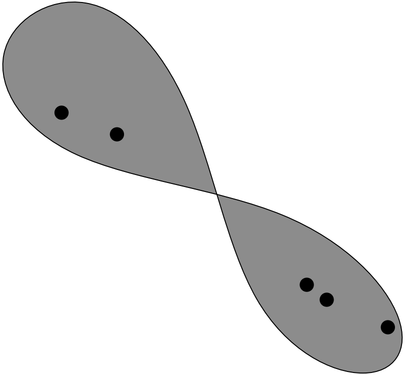
Nested Sampling

Graphical aid

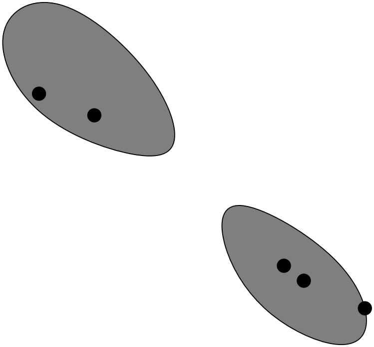


Nested Sampling

Graphical aid

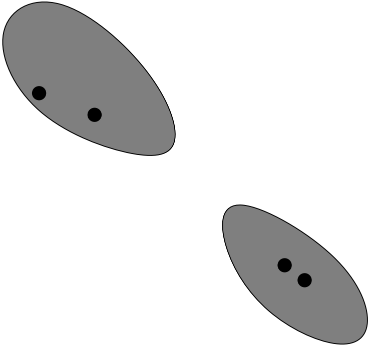


Nested Sampling
Graphical aid

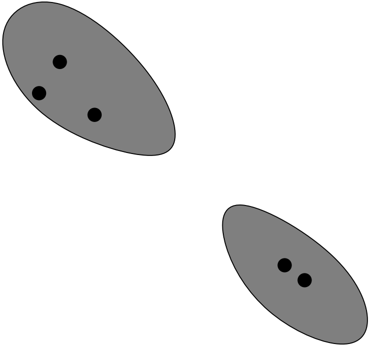


Nested Sampling

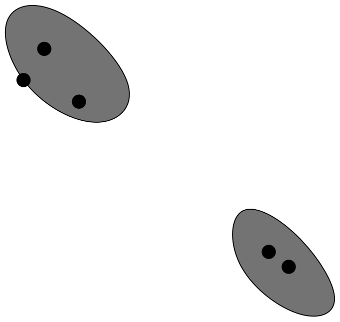
Graphical aid



Nested Sampling
Graphical aid

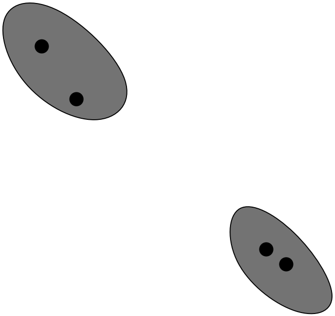


Graphical aid



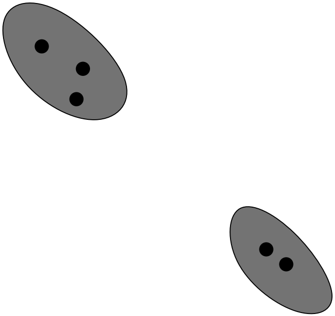
Nested Sampling

Graphical aid



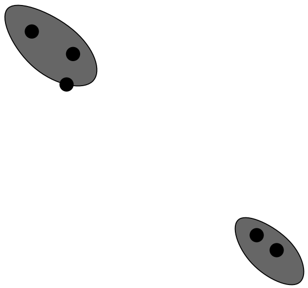
Nested Sampling

Graphical aid



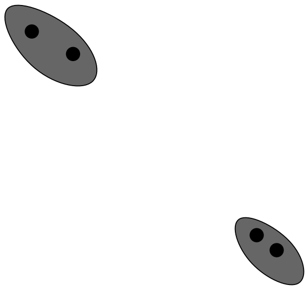
Nested Sampling

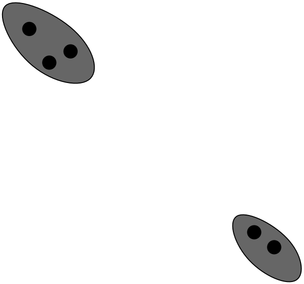
Graphical aid

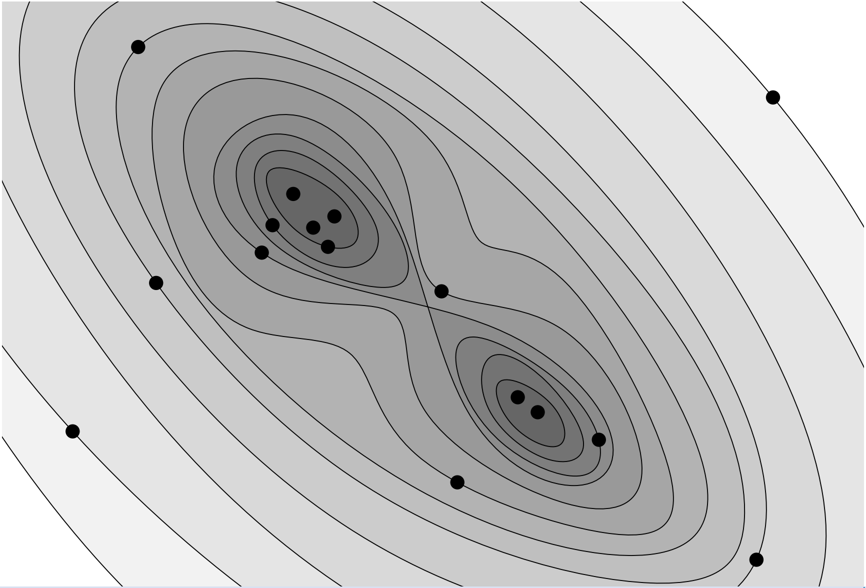


Nested Sampling

Graphical aid

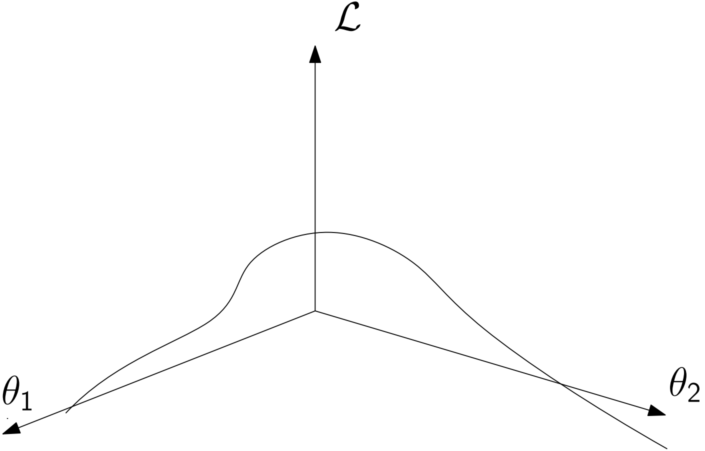






Nested sampling

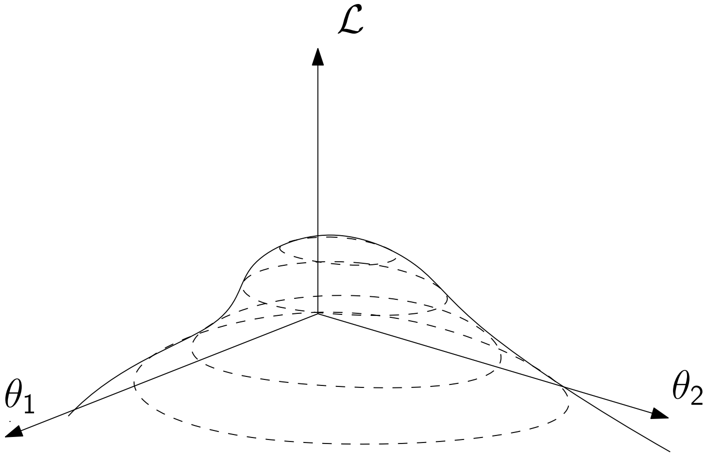
- The set of dead points are posterior samples with an appropriate weighting factor
- They can also be used to calculate evidences, since it sequentially updates the priors.
- The current set of live points is useful for performing clustering and constructing new proposed points.
- Algorithm terminates when prior has been compressed onto (and past) the posterior bulk (typical set).

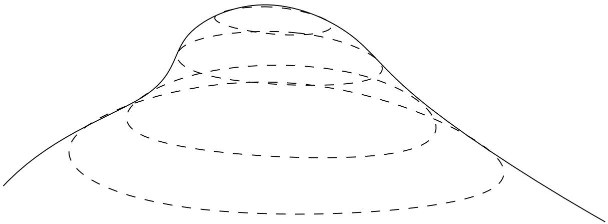


$\mathcal{L}$
θ_1
θ_2

Nested Sampling

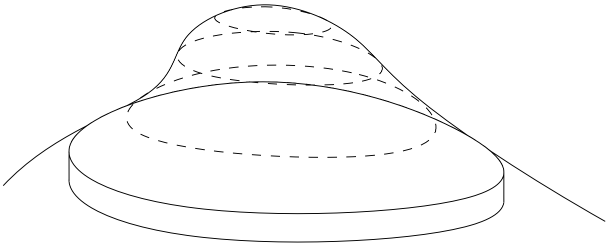
Calculating evidences

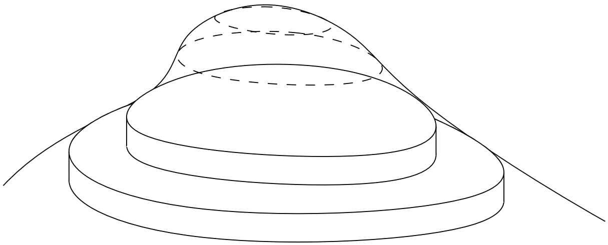


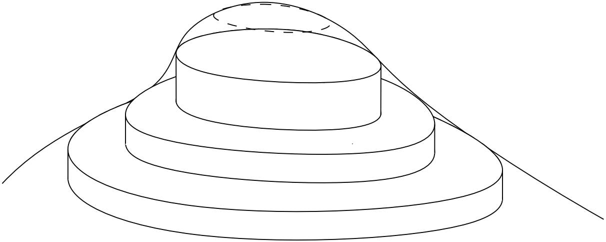


Nested Sampling

Calculating evidences

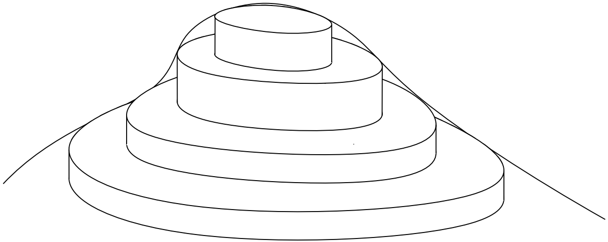


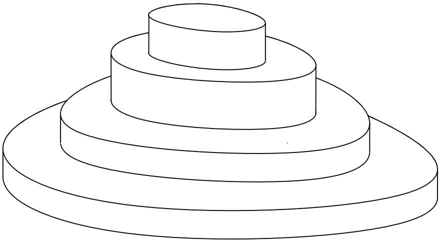


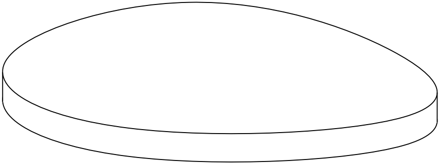


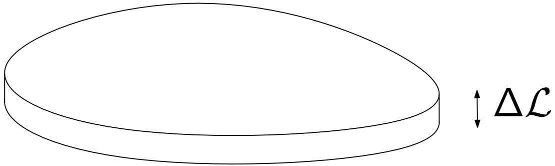
Nested Sampling

Calculating evidences

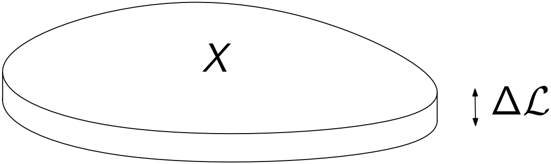


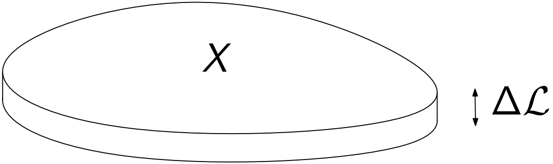




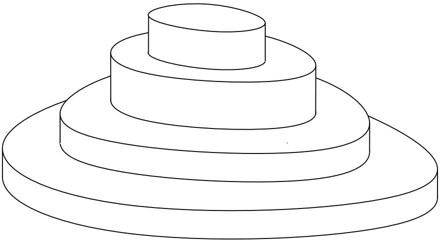

$\Delta\mathcal{L}$

Nested Sampling
Calculating evidences

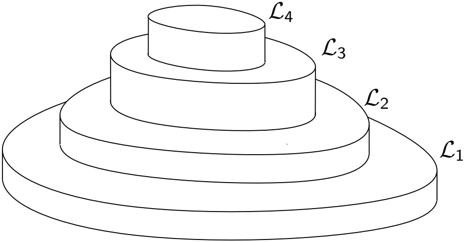




Volume $= X\Delta\mathcal{L}$

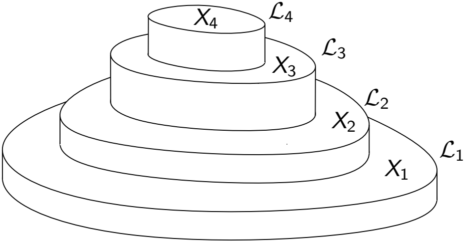


Nested Sampling
Calculating evidences

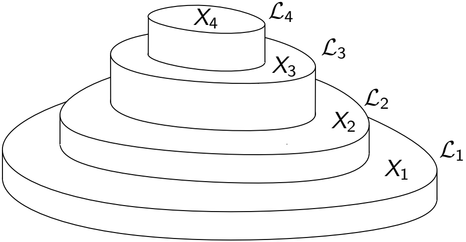


Nested Sampling

Calculating evidences



$$\mathcal{Z} \approx \sum_i X_i \Delta\mathcal{L}_i$$



- At each iteration, the likelihood contour will shrink in volume by $\approx 1/n$.
- Nested sampling zooms in to the peak of the posterior *exponentially*.

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i, \qquad X_{i+1} \approx \frac{n}{n+1} X_i, \qquad X_0 = 1$$

- Although this is only approximate, we can quantify the error

$$P(X_{i+1}|X_i) = \frac{1}{nX_i}\left(\frac{X_{i+1}}{X_i}\right)^{n-1}[0 < X_i < X_{i+1}]$$

- Integral can be expressed in one of two ways

$$\mathcal{Z} \approx \sum_i \Delta\mathcal{L}_i X_i = \sum_i \mathcal{L}_i \Delta X_i$$

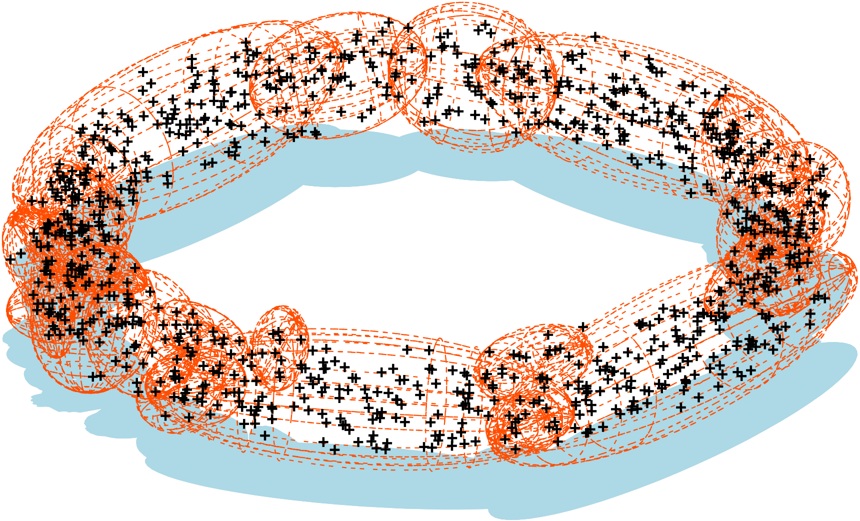
Sampling from a hard likelihood constraint

> *"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.
- `https://projecteuclid.org/euclid.ba/1340370944`

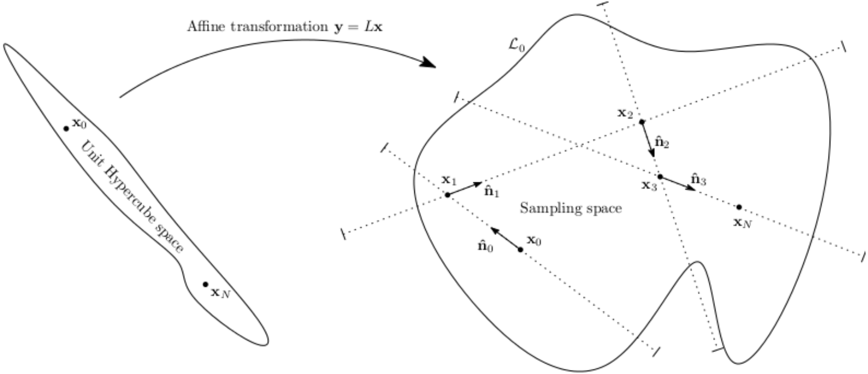
MultiNest

arXiv:0809.3437 arXiv:0704.3704 arXiv:1306.2144



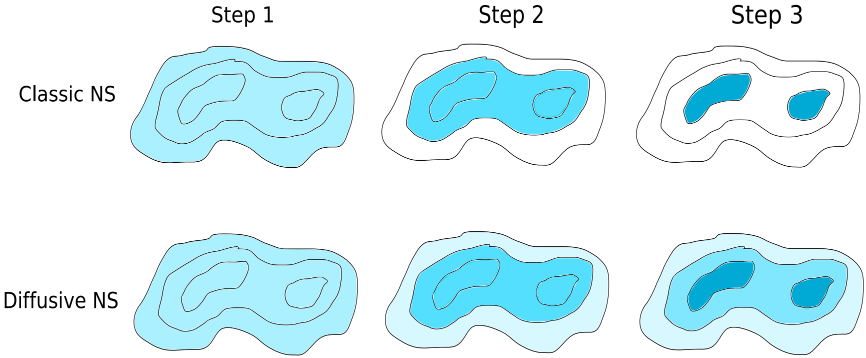
PolyChord

arXiv:1502.01856 arXiv:1506.00171



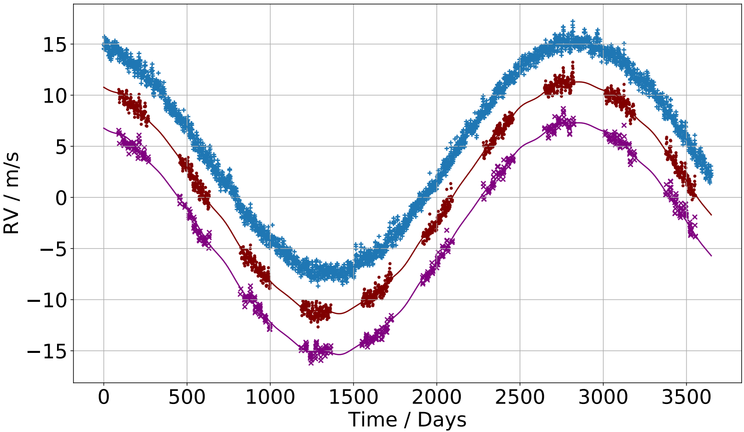
Diffusive nested sampling

arXiv:0912.2380



PolyChord vs MultiNest

- MultiNest excels in low dimensions $D < 10 - 20$.
- PolyChord can go up to ~ 150.
- Crossover is problem dependent
- PolyChord can also exploit fast-slow hierarchy


RV / m/s
Time / Days
15
10
5
0
−5
−10
−15
0
500
1000
1500
2000
2500
3000
3500

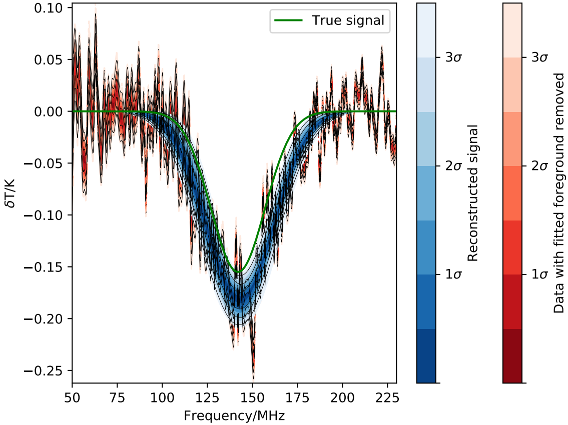
- Simple radial velocity model

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

- Fit each model to data.
- Posteriors on model parameters $[(K_p, \omega_p, \phi_p), p = 1 \cdots N]$ quantify knowledge of system characteristics.
- Evidences of models determine relative likelihood of number of planets in system
- This is an application where phase transitions matter

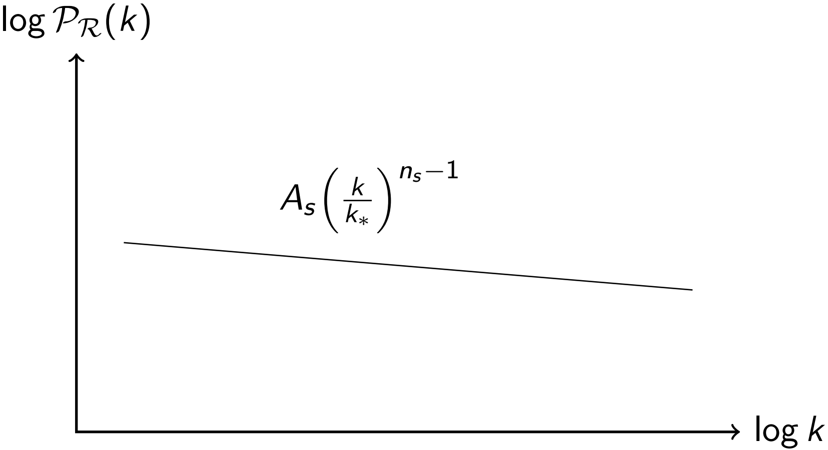
Nested sampling in action (Paper coming soon, Anstey, de Lera Acedo & Handley)

- Search for signal $T = T_{\mathrm{fg}} + T_{\mathrm{21cm}}$
- Fit parameterised models with/without T_{21cm}
- Compare evidences for signal detection
- Use evidences to quantify complexity of beam/sky models



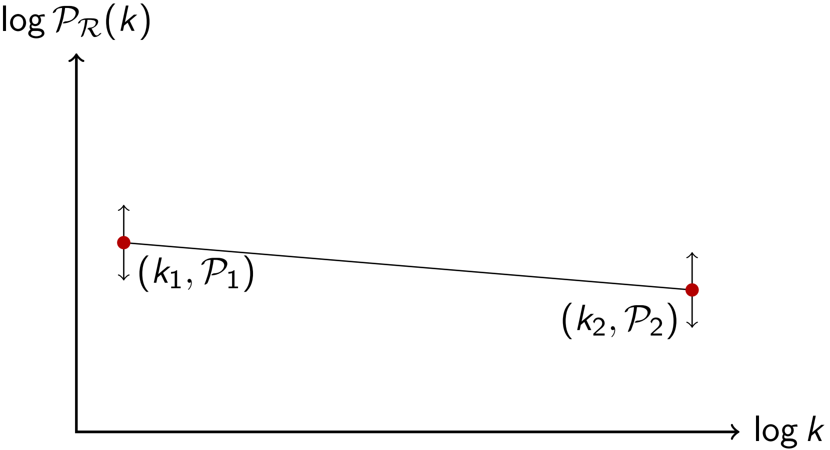
Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction

Nested Sampling in action (arXiv:1908.00906)



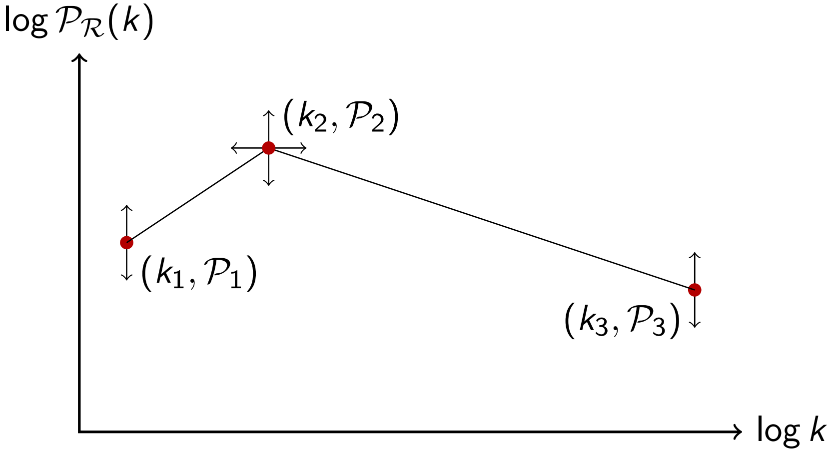
Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction

Nested Sampling in action (arXiv:1908.00906)



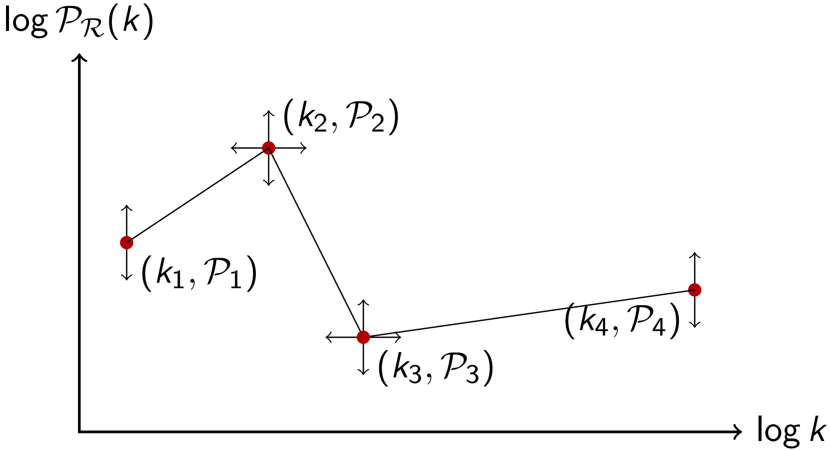
Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction

Nested Sampling in action (arXiv:1908.00906)



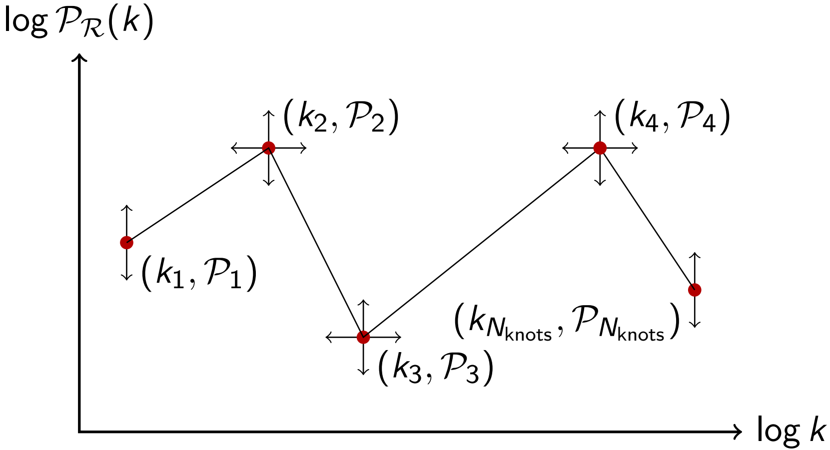
Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction

Nested Sampling in action (arXiv:1908.00906)

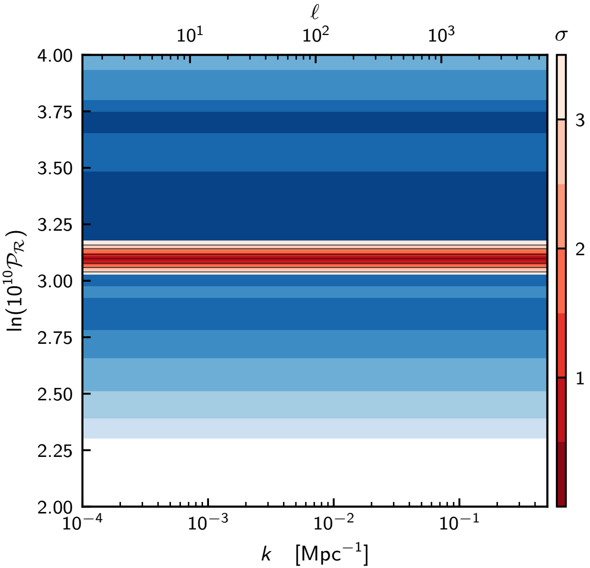


Primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ reconstruction

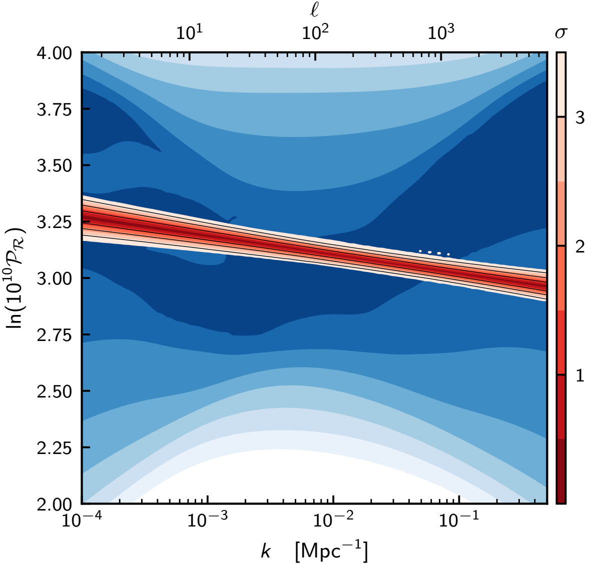
Nested Sampling in action (arXiv:1908.00906)



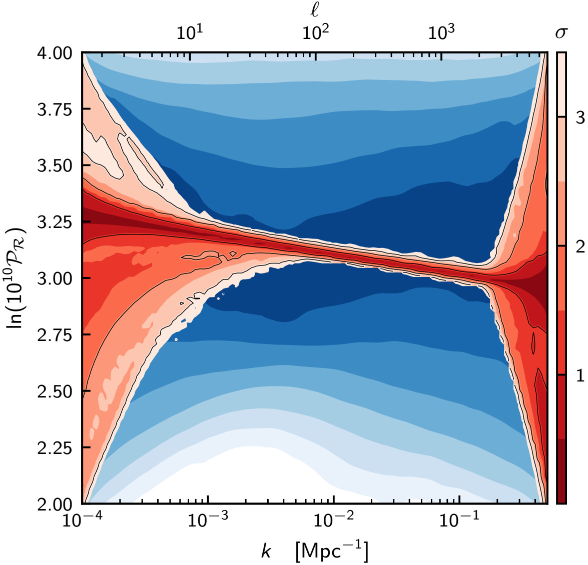
0 internal knots



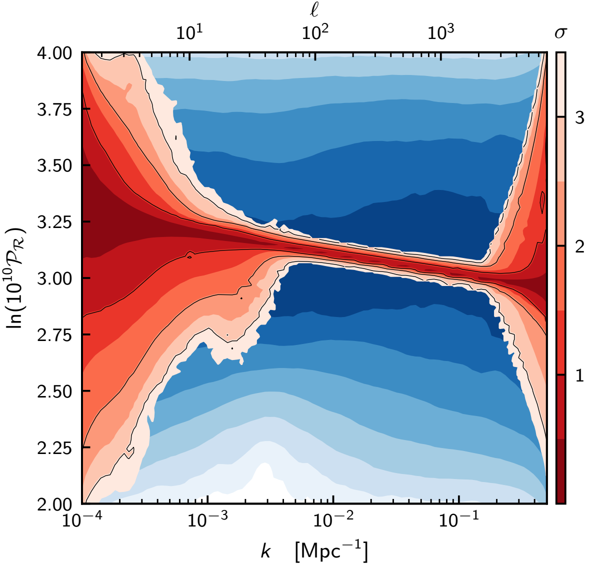
1 internal knot



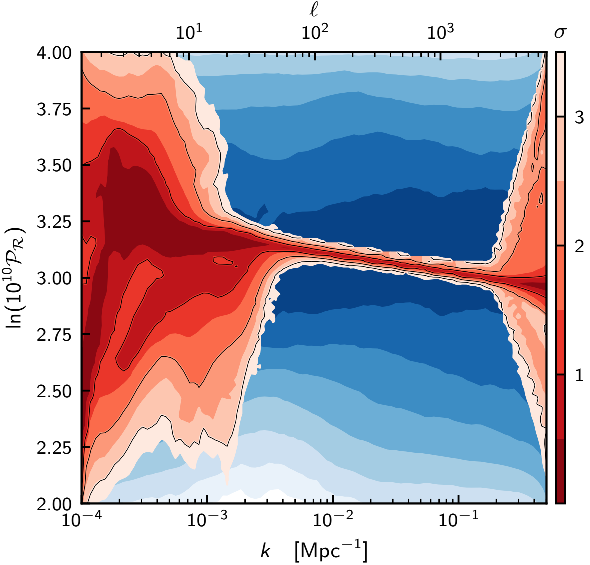
2 internal knots



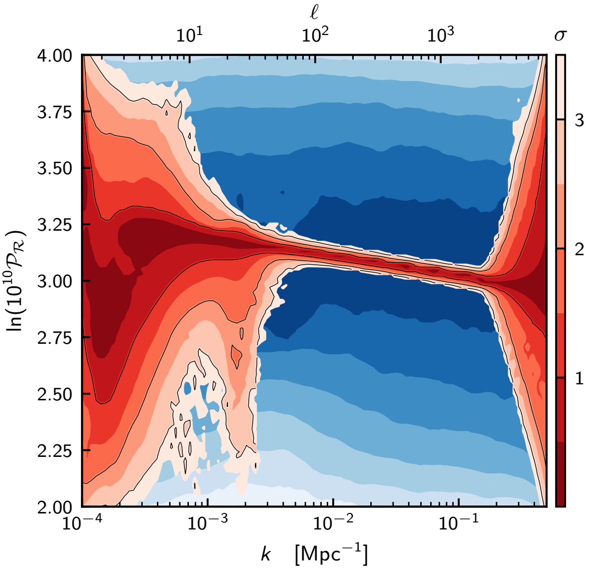
3 internal knots



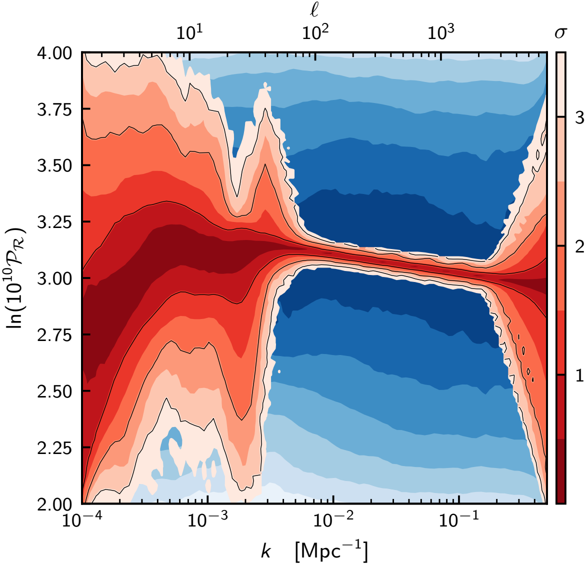
4 internal knots



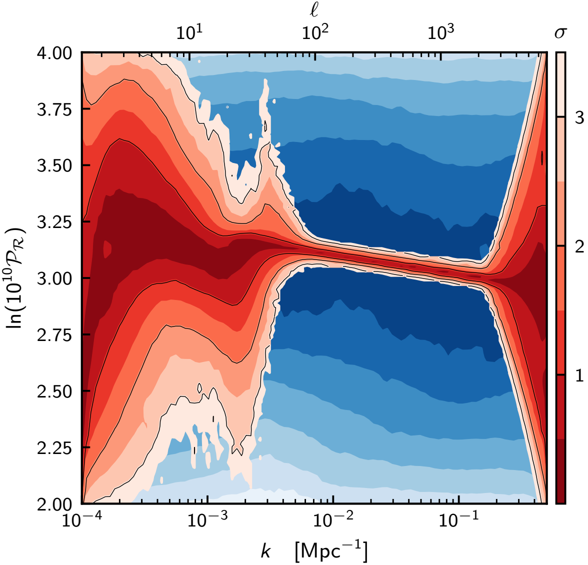
5 internal knots



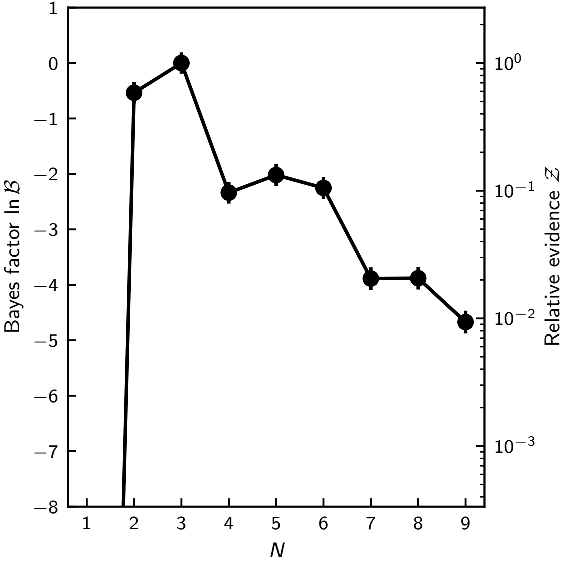
6 internal knots



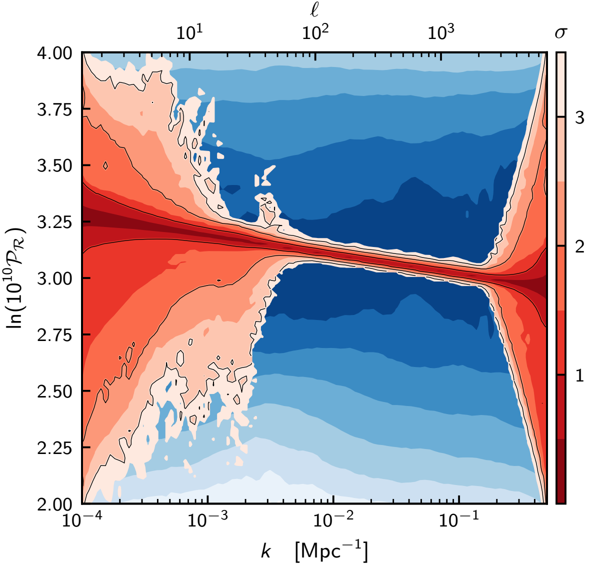
7 internal knots



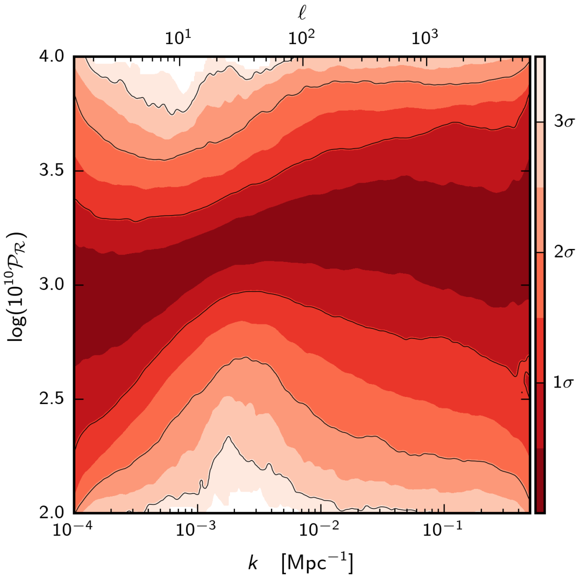
Bayes Factors



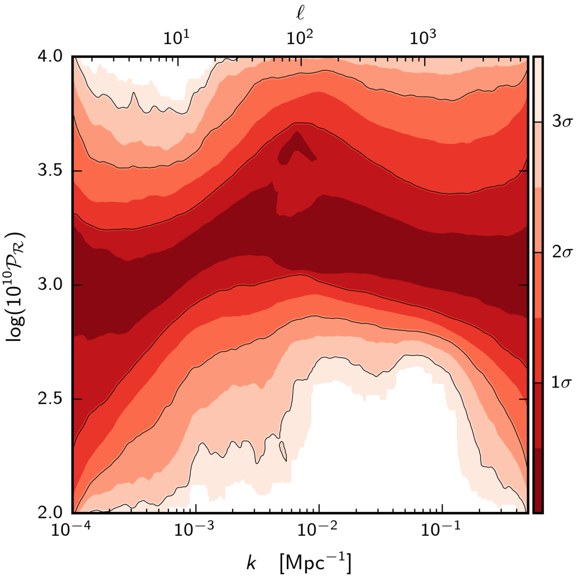
Marginalised plot



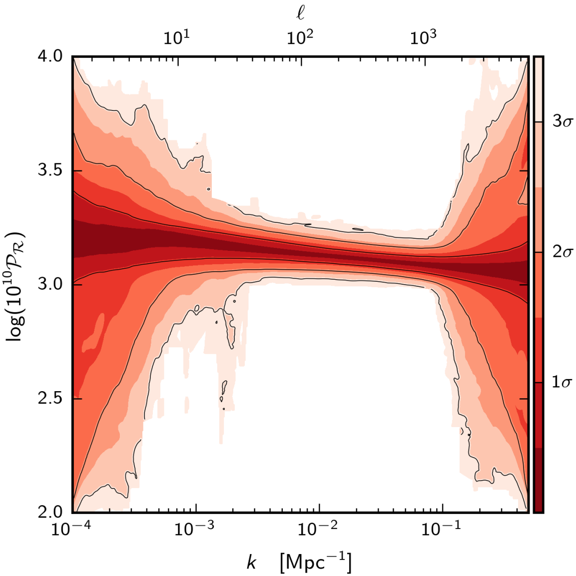
COBE (pre-2002)



COBE et al (2002)

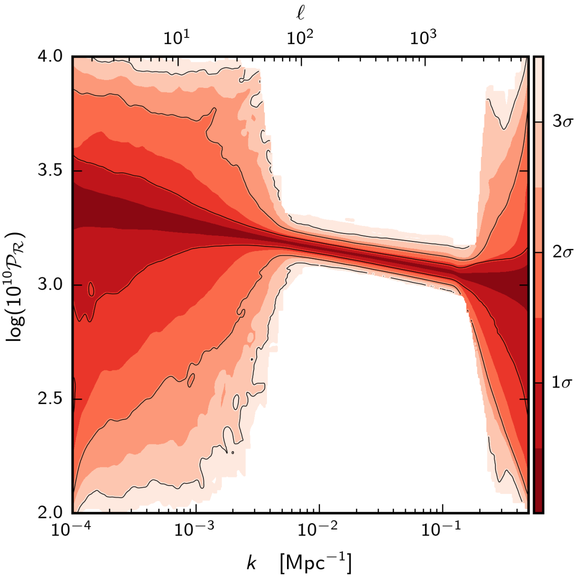


WMAP (2012)

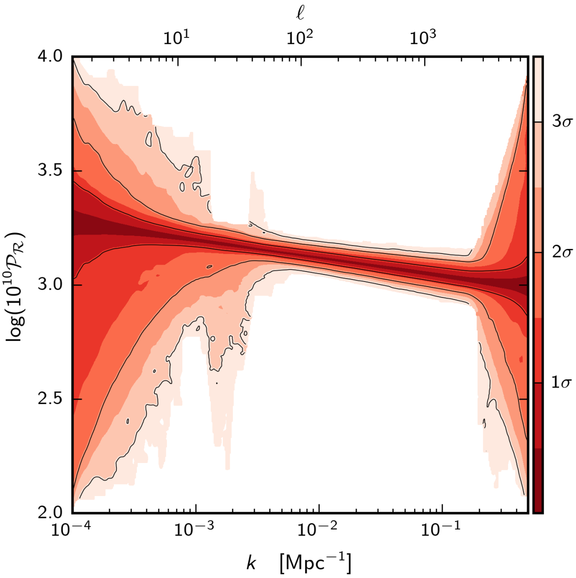


ℓ
10^1
10^2
10^3
4.0
3.5
3.0
2.5
2.0
$\log(10^{10}\mathcal{P}_\mathcal{R})$
10^{-4}
10^{-3}
10^{-2}
10^{-1}
$k \quad [\mathrm{Mpc}^{-1}]$
3σ
2σ
1σ

Planck (2013)

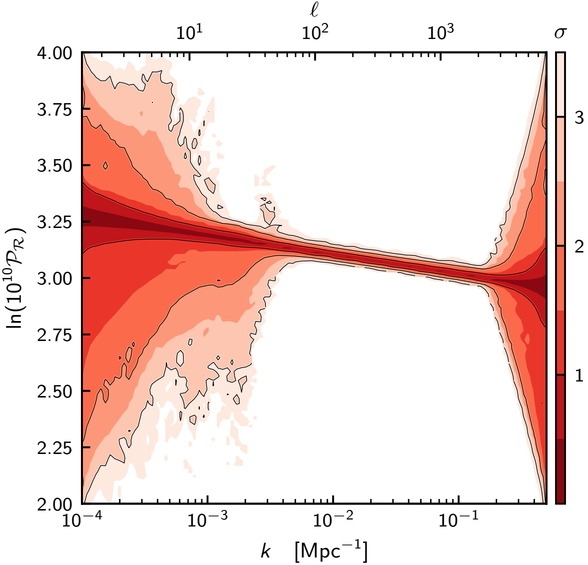


ℓ
10^1
10^2
10^3
4.0
3.5
3.0
2.5
2.0
$\log(10^{10}\mathcal{P}_\mathcal{R})$
10^{-4}
10^{-3}
10^{-2}
10^{-1}
k [Mpc^{-1}]
3σ
2σ
1σ



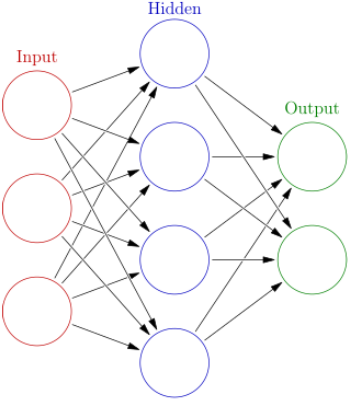
ℓ
10^1
10^2
10^3
4.0
3.5
3.0
2.5
2.0
$\log(10^{10}\mathcal{P}_\mathcal{R})$
10^{-4}
10^{-3}
10^{-2}
10^{-1}
k [Mpc^{-1}]
3σ
2σ
1σ

Planck (2018)



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

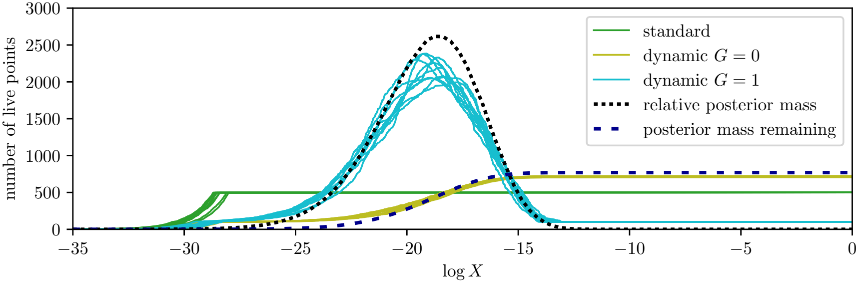
- Neural networks require:
  - Training to find weights
  - Choice of architecture/topology
- Bayesian NNs treat training as a model fitting problem
- Compute posterior of weights (parameter estimation)
- Use evidence to determine best architecture (model comparison)
- Paper coming soon (Javid, Handley, Lasenby & Hobson)
  - Bayesian evidences correlate with out-of-sample performance
  - Can be used to determine width and number of hidden layers



Unweaving runs

- John Skilling noted that two nested sampling runs can be combined in likelihood order to produce a valid run with a larger number of live points.
- The reverse is also true (arXiv:1704.03459).
- In general, a run with n live points can be "unweaved" into n runs with a single live point.
- Useful for providing convergence diagnostics and better parameter estimation (arXiv:1804.06406).

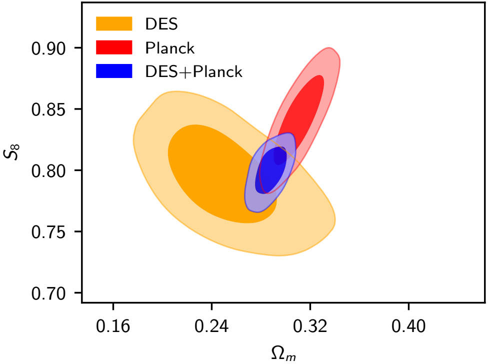
Advances in nested sampling (arXiv:1704.03459, dynesty: arXiv:1904.02180)



The number of live points can be varied dynamically in order to oversample regions of interest

Other uses of nested sampling

- Nested sampling estimates the density of states ΔX_i, and hence gives you access to a lot more than just posterior samples
- Kullback-Liebler divergence (arXiv:1607.00270)
- Bayesian model dimensionality (arXiv:1903.06682)
- Suspiciousness & Tension quantification (arXiv:1902.04029)



Multi-temperature sampling

- By compressing from prior to posterior, Nested Sampling's weighted samples are fundamentally different from traditional MCMC.
- Nested sampling tails and peaks equally.
- We can define the "temperature" of a distribution in analogy with thermodnyamics:

$$\log L \sim E \Rightarrow P \propto e^{-\beta E} = e^{-E/kT}, \quad \beta = 1$$

- Sampling at different temperatures can be useful for exploring tails.
- Nested sampling runs give you the full partition function

$$\log Z(\beta) \approx \sum_i \mathcal{L}_i^\beta \Delta X_i$$

Nested importance sampling

Future research

- Much of the time spent in a nested sampling run is spent "compressing the tails".
- Posterior-repartitioned nested sampling gives one way of speeding this up (arXiv:1908.04655)
- Sometimes we have a-priori good knowledge of the posterior bulk (analagous to an MCMC proposal distribution).

$$Z_0 = \int L(\theta)\pi_0(\theta)d\theta, \qquad Z_1 = \int L(\theta)\pi_1(\theta)d\theta$$
$$= \int L(\theta)\pi_1(\theta)\frac{\pi_0(\theta)}{\pi_1(\theta)}d\theta = \left\langle \frac{\pi_0(\theta)}{\pi_1(\theta)} \right\rangle_{P_1}$$

- This importance weighting only works if you have a lot of tail samples.

N-σ contours

Future research

- Traditional posterior samples only allow you to plot contours out to 2-3σ.
- Nested sampling fully samples the tails, so in theory one could do 20σ contours.
- Requires further thought in alternatives to kernel density estimation.

Things every nested sampling user should know

- "Burn in" can take a while, and results are not informative until then.
- Reducing the stopping criterion does not appreciably change run-time, but does reduce reliability.
- Run time is linear in the number of live points, so reduce this for exploratory runs $\sim \mathcal{O}(10)$, but increase to $\sim \mathcal{O}(1000)$ for production-ready runs.
- Most nested sampling algorithms are intensely parallelisable, and work best in pure MPI mode (no openMP).

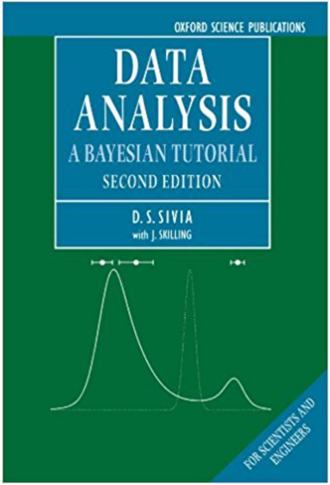
Key software

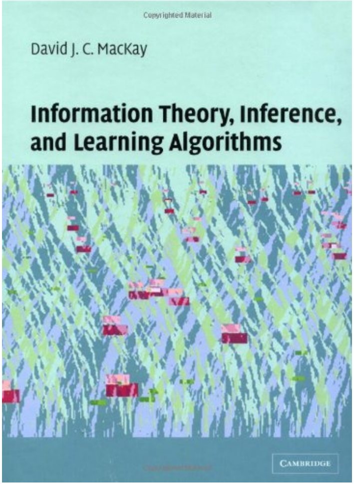
MultiNest	`github.com/farhanferoz/MultiNest`
PolyChord	`github.com/PolyChord/PolyChordLite`
DNest	`github.com/eggplantbren/DNest3`
dynesty	`github.com/joshspeagle/dynesty`
anesthetic	nested sampling visualisation `github.com/williamjameshandley/anesthetic`
fgivenx	posterior plotting of functions `github.com/williamjameshandley/fgivenx`
cosmology	Implemented as an alternative sampler in CosmoMC, MontePython, cosmosis, cobaya & GAMBIT

Summary

- Nested sampling is a rich framework for performing the full pipeline of Bayesian inference
- Plenty of further work to do on the underlying theory
  - (If any students/postdocs are interested, I have a large stack of projects waiting to be explored)
- Some understanding is required in order to operate & get the most from nested sampling chains.

Further reading





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