Nested Sampling
Bayesian Inference

- Two stages:
  1) Parameter estimation - Posterior
  2) Model selection - Evidence

\[
P(d|\theta, M) \times P(\theta|M) = P(\theta|d, M) \times P(d|M)
\]

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood</td>
<td>Prior</td>
</tr>
<tr>
<td></td>
<td>Posterior</td>
</tr>
<tr>
<td></td>
<td>Evidence</td>
</tr>
</tbody>
</table>
Model Selection

• Apply Bayes theorem to models rather than parameters

\[
P(M_i|D) = \frac{P(D|M_i)P(M_i)}{P(D)},
\]

• The normalisation here can be written

\[
P(D) = \sum_j Z_j \pi_j \quad Z_i \equiv P(D|M_i) = \int P(D|\theta, M_i)P(\theta|M_i)d\theta.
\]

• So that the posterior of the model can be written in terms of the evidences and priors for the models

\[
P(M_i|D) = \frac{Z_i \pi_i}{\sum_j Z_j \pi_j}.
\]
Model Selection

• For uniform priors on the models, we prefer a model with a larger Evidence

\[
\frac{P(M_1|D)}{P(M_2|D)} = \frac{\mathcal{Z}_1 \pi_1}{\mathcal{Z}_2 \pi_2}
\]

• Evidence is key for Bayesian model selection!
• How can we calculate the evidence?
The evidence

- Evidence is integral over likelihood and prior

\[ P(\mathcal{D}|\mathcal{M}) \equiv \mathcal{Z} = \int P(\mathcal{D}|\theta, \mathcal{M})P(\theta|\mathcal{M})d\theta. \]

\[ \mathcal{Z} = \int \mathcal{L}(\theta)\pi(\theta)d\theta. \]

- Typically the integral is in a high-dimensional space, but only a small region contributes significantly to integral. Need to find it!
Limitations of MCMC

- MCMC with Metropolis-Hastings typically focuses on peak of posterior and explores in that vicinity.
- Low sampling in tails of distribution. Not a problem for parameter estimation, but can be when calculating evidence.
- Difficult to handle multimodal posterior distributions - may get trapped.
Multimodal likelihood

- MH may get trapped in local maximum without exploring full likelihood shape

ICIC
Nested Sampling

• Goal of efficiently evaluating evidence and returning posterior estimate.
Nested Sampling

• Imagine ordering set of likelihood points
• Introduces *prior volume*: fraction of prior contained within an iso-likelihood contour

\[ X(L) = \int_{L(\theta) > L} \pi(\theta) d\theta \]
• Use to transform evidence calculation from multidimensional integral to a 1D integral

\[ Z = \int \mathcal{L}(\theta)\pi(\theta)d\theta. \quad \rightarrow \quad Z = \int_0^1 \mathcal{L}(X)dX. \]

• Ordered L(X) then gives evidence via 1D integration e.g. via quadrature

\[ Z = \sum_{i=1}^{M} \mathcal{L}_i w_i \quad \quad \quad w_i = \frac{1}{2}(X_{i-1}^--X_{i+1}) \]

• Points chosen randomly from region L(X) are representative of posterior

\[ P(X_i) = \frac{\mathcal{L}(X_i)w_i}{Z} \]
Nested Sampling

- Uniformly sample from prior maintaining a population of live points that is updated so that they contract around the peak(s) of posterior

- Assign X values on basis of statistics of uniform dist

\[
\log X_i \approx -\left( i \pm \sqrt{i} \right) / N
\]

Points exponentially hone in on high \( L(X) \) as \( X_k \sim \exp(-k/n) \) for \( n \) points
five live points chosen uniformly from prior
Define likelihood contour from lowest point
Delete that point (storing it’s value)
Select a new point uniformly sampled subject to requirement $L(X_{\text{new}}) > L(X_{\text{old}})$.
Iterate - contour shrinks by $X \sim \exp(1/n)$
Successive points map likelihood in an ordered way.
Stopping criteria

- Can decide to stop based on error estimate on evidence
- Likelihood increases, but separation of points decreases, so contribution to integral converges

\[ \Delta Z_i = L_{\text{max}} X_i \]
Ultimately left with a set of points with known $\{\theta_i, L_i\}$, inferred $X_i$ and estimate of evidence $Z$.
Incomplete algorithm

- Uniform sampling from prior subject to \( L(X) > \lambda \) is not straightforward.
- e.g. Ellipsoidal rejection sampling (MultiNest)

**Figure 2.** Cartoon of ellipsoidal nested sampling from a simple bimodal distribution. In (a) we see that the ellipsoid represents a good bound to the active region. In (b)–(d), as we nest inwards we can see that the acceptance rate will rapidly decrease as the bound steadily worsens. (e) illustrates the increase in efficiency obtained by sampling from each clustered region separately.
Codes

• Two examples:
  MultiNest - ellipsoidal rejection sampling
  PolyChord - slice sampling

( python wrapper - pymultinest)
likelihood functions are usually quite benign, often resembling a
mological parameter estimation and model selection, for which the
problems. More efficient (and certainly no less so) than standard Metropolis–
challenging, and it is important to demonstrate that M
Trotta et al. 2008). None the less, not all likelihood functions are as
McLachlan 2003; Feroz, Marshall & Hobson 2008a; Feroz et al.

ESTIMATION AND MODEL SELECTION

An important area of inference in astrophysics is that of cos-

6.2 Toy model 2: Gaussian shells likelihood

high dimension.

Successive iterations of the M
(2009 The Authors. Journal compilation

Table 2. The number of likelihood evaluations and sampling

The local log-evidence values of each mode for the toy

Table 2.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Toy model 1</th>
<th>Toy model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter 1</td>
<td>1.75</td>
<td>0.08</td>
</tr>
<tr>
<td>Parameter 2</td>
<td>0.09</td>
<td>0.24</td>
</tr>
<tr>
<td>Parameter 3</td>
<td>0.09</td>
<td>0.19</td>
</tr>
<tr>
<td>Parameter 4</td>
<td>0.11</td>
<td>0.27</td>
</tr>
<tr>
<td>Parameter 5</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Parameter 6</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Parameter 7</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>

From FH08

Like

ULTI

ULTI

ULTI

ULTI

ULTI

ULTI
Examples
LISA Gravitational Wave detection
LISA time series analysis

• Cosmic strings produce beamed burst of gravitational waves via cusp formation

\[
|h(f)| = \begin{cases} 
Af^{-4/3} & f < f_{\text{max}} \\
Af^{-4/3} \exp\left(1 - \frac{f}{f_{\text{max}}}\right) & f > f_{\text{max}}
\end{cases}
\]

phase by \exp(2\pi if t_c),

Ferhoz, Gair + 2010
MultiNest used to search mock LISA timestream for cosmic string signal
degeneracies in parameters e.g. antipodal sky locations

![Graph of marginalized posteriors for various parameters](image)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Marginalized Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colatitude</td>
<td></td>
</tr>
<tr>
<td>Longitude</td>
<td></td>
</tr>
<tr>
<td>Burst time</td>
<td></td>
</tr>
<tr>
<td>Burst amplitude</td>
<td></td>
</tr>
<tr>
<td>Break frequency</td>
<td></td>
</tr>
<tr>
<td>Waveform polarization</td>
<td></td>
</tr>
</tbody>
</table>

ICIC: 
- colatitude
- longitude
- burst time
- burst amplitude
- break frequency
- waveform polarization
• Model selection to determine type of burst: cosmic string versus Sine-Gaussian model
Bayesian Image detection
Bayesian Image detection

- e.g. finding circular objects in noisy data

\[
\tau(x; a) = A \exp \left[ -\frac{(x - X)^2 + (y - Y)^2}{2R^2} \right] \quad a = \{X, Y, A, R\}
\]


- Sky model will be sum of such objects

\[
D = n + \sum_{k=1}^{N_{\text{obj}}} s(a_k),
\]

- Gaussian noise determines likelihood

\[
\Pr(D \mid \theta) = \frac{\exp \left\{ -\frac{1}{2} [D - s(a)]'N^{-1}[D - s(a)] \right\}}{(2\pi)^{N_{\text{pix}}/2} |N|^{1/2}},
\]

Assume prior is separable by object

\[
\Pr(\theta) = \Pr(N_{\text{obj}})\Pr(a) = \Pr(N_{\text{obj}})\Pr(a_1)\Pr(a_2) \cdots \Pr(a_{N_{\text{obj}}})
\]
Figure 1. The toy problem discussed in Section 4.3. The 200 × 200 pixel test image (left panel) contains eight discrete Gaussian-shaped objects of varying widths and amplitudes; the parameters $X_k$, $Y_k$, $A_k$ and $r_k$ for each object are listed in Table 1. The corresponding data map (right panel) has independent Gaussian pixel noise added with an rms of 2 units. This figure is available in colour in the on-line version of the journal on Synergy.
N=1 fitting

- Simplified analysis with just one object: leads to multimodal posterior with peaks at object locations

**Figure 3.** The two-dimensional conditional log-posterior distributions in the \((X, Y)\)-subspace for the toy problem illustrated in Fig. 1, where the model contains a single object parametrized by \(a = \{X, Y, A, R\}\). The values of the amplitude \(A\) and size \(R\) are conditioned at \(A = 0.75, R = 5\) (left panel) and \(A = 0.75, R = 10\) (right panel). This figure is available in colour in the on-line version of the journal on Synergy.
Identifying real sources

• Not all peaks in posterior will be sources - apply model selection to distinguish

\[ H_0 = \text{`a cluster with } M_{g,\text{min}} < M_g \leq M_{g,\text{lim}} \text{ is centred in } S', \text{ no-cluster} \]

\[ H_1 = \text{`a cluster with } M_{g,\text{lim}} < M_g < M_{g,\text{max}} \text{ is centred in } S', \text{ cluster} \]

SZ cluster detection with AMI Feroz+ 2009

ICIC
Odds ratio \( e^{12.2 \pm 0.2} \quad 0.32 \pm 0.03 \)
Global particle physics analysis

• Combine particle physics and astrophysics constraints to learn about beyond the Standard Model physics

<table>
<thead>
<tr>
<th>Observable</th>
<th>Mean value</th>
<th>Uncertainties</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_W$ [GeV]</td>
<td>80.399 0.023 0.015</td>
<td></td>
<td>[34]</td>
</tr>
<tr>
<td>$\sin^2 \theta_{eff}$</td>
<td>0.2315 0.0001</td>
<td>0.00015</td>
<td></td>
</tr>
<tr>
<td>$\delta a^SUSY_\mu \times 10^{10}$</td>
<td>28.7 8.0</td>
<td>2.0</td>
<td>[35]</td>
</tr>
<tr>
<td>$BR(\bar{B} \rightarrow X_s \gamma) \times 10^4$</td>
<td>3.55 0.26</td>
<td>0.30</td>
<td>[36]</td>
</tr>
<tr>
<td>$R_{\Delta M_{B_s}}$</td>
<td>1.04 0.11</td>
<td>-</td>
<td>[37]</td>
</tr>
<tr>
<td>$\frac{BR(\bar{B}<em>s \rightarrow \tau \nu)}{BR(B_u \rightarrow \tau \nu)</em>{SM}} \times 10^2$</td>
<td>1.63 0.54</td>
<td>-</td>
<td>[36]</td>
</tr>
<tr>
<td>$\Delta_0 \times 10^2$</td>
<td>3.1 2.3</td>
<td>-</td>
<td>[38]</td>
</tr>
<tr>
<td>$BR(B \rightarrow D_{\tau \nu}) \times 10^2$</td>
<td>41.6 12.8</td>
<td>3.5</td>
<td>[39]</td>
</tr>
<tr>
<td>$R_{123}$</td>
<td>0.999 0.007</td>
<td>-</td>
<td>[40]</td>
</tr>
<tr>
<td>$BR(D_s \rightarrow \tau \nu) \times 10^2$</td>
<td>5.38 0.32</td>
<td>0.2</td>
<td>[36]</td>
</tr>
<tr>
<td>$BR(D_s \rightarrow \mu \nu) \times 10^3$</td>
<td>5.81 0.43</td>
<td>0.2</td>
<td>[36]</td>
</tr>
<tr>
<td>$BR(D \rightarrow \mu \nu) \times 10^4$</td>
<td>3.82 0.33</td>
<td>0.2</td>
<td>[36]</td>
</tr>
<tr>
<td>$\Omega \chi^2$</td>
<td>0.1109 0.0056</td>
<td>0.012</td>
<td>[41]</td>
</tr>
<tr>
<td>$m_h$ [GeV]</td>
<td>125.8 0.6</td>
<td>2.0</td>
<td>[19]</td>
</tr>
<tr>
<td>$BR(\bar{B}_s \rightarrow \mu^+ \mu^-)$</td>
<td>$3.2 \times 10^{-9}$ 1.5 $\times 10^{-9}$</td>
<td>10%</td>
<td>[20]</td>
</tr>
<tr>
<td>Limit (95% CL)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tau$ (theor.)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sparticle masses

| $m_0, m_{1/2}$ | ATLAS, $\sqrt{s} = 8$ TeV, 5.8 fb$^{-1}$ 2012 limits | [17] |
| $m_A, \tan \beta$ | CMS, $\sqrt{s} = 7$ TeV, 4.7 fb$^{-1}$ 2012 limits | [18] |
| $m_\chi - \sigma_{SI}^X_{1\%}$ | XENON100 2012 limits (224.6 $\times$ 34 kg days) | [21] |
Constrained MSSM

- WMAP7+ATLAS+CMS Higgs mass + XENON100+...

Strege+ (2013)  
arXiv:1212.2636
Other useful packages
Affine invariance

• Common problem for MCMC is need to sample likelihood with narrow degeneracies
• Requires carefully chosen proposal distribution
• Would like to exploit an affine transformation to a more symmetric space

\[
p(x) \propto f\left(-\frac{(x_1 - x_2)^2}{2\epsilon} - \frac{(x_1 + x_2)^2}{2}\right)
\]

\[
y_1 = \frac{x_1 - x_2}{\sqrt{\epsilon}}, \quad y_2 = x_1 + x_2.
\]
Emcee - “The MC Hammer”

- Emcee - affine invariant ensemble sampler (available as python module)  
  Foreman-Mackey+ (2012)
- Evolve ensemble of walkers (c.f. Metropolis-Hastings)
- Stretch-move:

For walker at $X_k$, chose another walker $X_j$ and propose move

$$X_k(t) \rightarrow Y = X_j + Z[X_k(t) - X_j],$$

$Z$ random variable drawn from distribution chosen to ensure detailed balance

$$g(z) \propto \begin{cases} 
\frac{1}{\sqrt{Z}} & \text{if } z \in \left[\frac{1}{a}, a\right] \\
0 & \text{otherwise}
\end{cases}$$

Accept proposal according to

$$q = \min \left(1, Z^{N-1} \frac{p(Y)}{p(X_k(t))} \right),$$

Repeat for all walkers in series.
• Versatile module for general MCMC problems
COSMOSIS

- Modular Cosmology analysis code
- includes MultiNest, CosmoMC
- integrates likelihoods from Planck, WMAP, DES,...
- easy to switch in and out different samplers/datasets
Conclusions

- Nested sampling offers alternative to MCMC to sampling posterior and provides evidence
- Nested sampling is well suited to problems with multi-modal posteriors
- Codes like MultiNest and PolyChord are freely available and should be in your toolkit
- Many other useful packages: COSMOSIS, EMCEE, SuperBayes, ...
Some notation

- Likelihood

\[ P(\mathcal{D}|\theta, \mathcal{M}) \equiv \mathcal{L}, \]

\[ P(\theta|\mathcal{M}) \equiv \pi, \]

\[ P(\mathcal{D}|\mathcal{M}) \equiv \mathcal{Z} = \int P(\mathcal{D}|\theta, \mathcal{M})P(\theta|\mathcal{M})d\theta. \]

\[ \mathcal{Z} = \int \mathcal{L}(\theta)\pi(\theta)d\theta. \]

\[ \mathcal{P} = \frac{\mathcal{L} \times \pi}{\mathcal{Z}}. \]