

# Entropic Priors and Bayesian Model Selection

Brendon J. Brewer  
Department of Physics  
University of California, Santa Barbara

<http://www.physics.ucsb.edu/~brewer/>

# The Motivation

- Model selection

$$\frac{P(H_2|D)}{P(H_1|D)} = \frac{P(H_2)}{P(H_1)} \times \frac{P(D|H_2)}{P(D|H_1)}$$

where

$$P(D|H_i) = Z_i = \int_{\theta_i} p(\theta_i|H_i)P(D|\theta_i, H_i) d\theta_i$$

- If all of these  $P$ 's model our prior knowledge well, we're done!

# Publish the Evidence!

- Good advice, people don't have to recompute the  $Z$  of other models if someone has already done it.
- People are doing this enthusiastically – in some fields of astronomy at least.
- But it's not the whole story...

# Sure Thing

- Lottery with 1,000,000 tickets
- Winner is  $D = 263878$

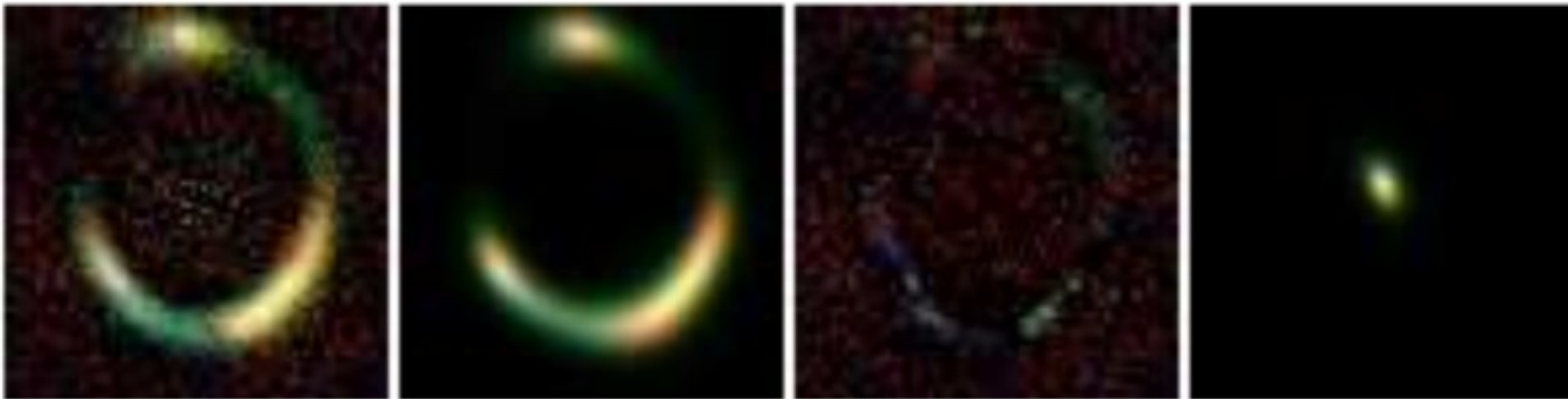
$H_1$  : Lottery is fair,  $P(D|H_1) = 10^{-6}$  for all  $D$ ,  
and hence for the observed  $D$ .

$H_2$  : Lottery is rigged towards ticket 263878,  
 $P(D|H_2) = 1$  for the observed  $D$ , so  $H_2$  wins  
on evidence by a factor of a million

Jaynes: You forgot the 999,999 other “sure things”. So  $P(H_2) = 0.5 \times 10^{-6} \ll 0.5$ .

# Another Example

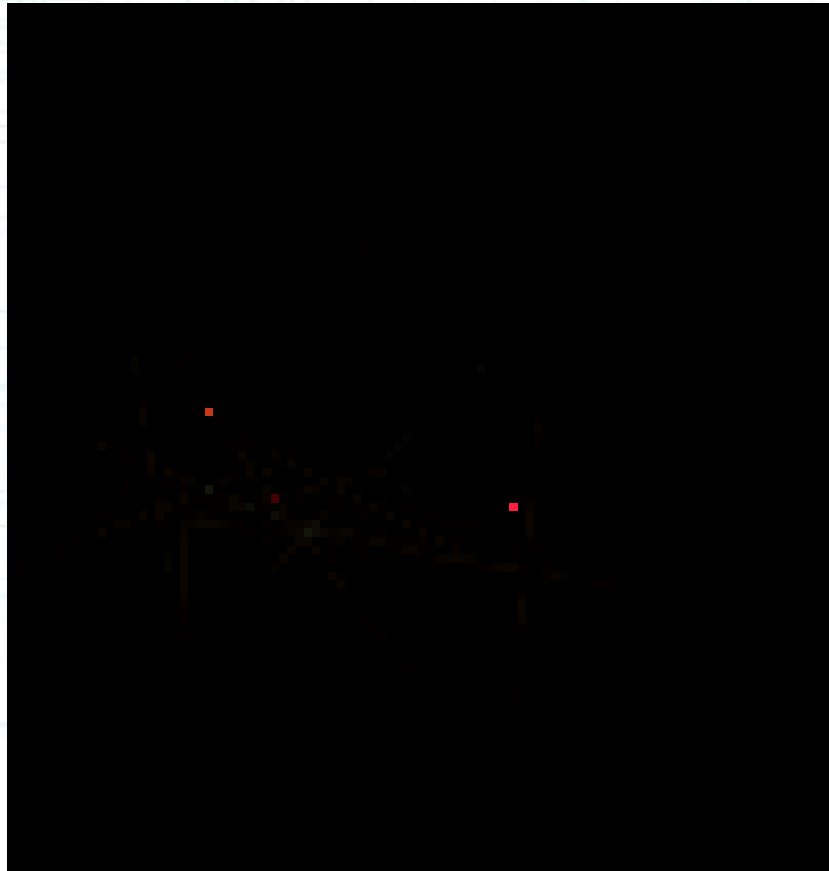
- Want to unlens gravitationally lensed galaxies. Get source light profile and projected lens mass profile simultaneously.



Example from Marshall, P. et al 2007.  
Used Sersic profile model for source

# I like complicated priors

- **Sorry Peter!**
- Inspired by, but not equivalent to, Radford Neal's Dirichlet Diffusion Trees.  
Brewer et al, in prep. Work with Phil Marshall, Tommaso Treu, Geraint Lewis and others.



# Cutting off your thumb with Occam's Razor

- Could calculate  $Z_{\text{sersic}}$  and  $Z_{\text{complex}}$  and “do model selection”
- This is only useful if
$$P(\text{Sersic}) = P(\text{Complex}) \sim 1/2$$
- This is silly. I'll bet you a million dollars that the actual galaxy profile is not in the Sersic family.
- We use it because of pragmatism and because its parameters reflect interesting summaries of the full, complex light profile. Not because it's a real theory.

# Moral

- Evidence doesn't “objectively rank models”.
- Conclusions also depend on prior probability.
- But we already knew that.

$$\frac{P(H_2|D)}{P(H_1|D)} = \frac{P(H_2)}{P(H_1)} \times \frac{P(D|H_2)}{P(D|H_1)}$$

# Guessing

- These two examples share a common theme
- Naive “model selection” favoured the model with sharper predictions
- But those models both happened to have tiny prior probability...
- This often makes sense: In general, the more specific a proposition is, the less probable.

$$P(A \text{ and } B \text{ and } C) \leq P(A \text{ and } B) \leq P(A)$$

# Can we be more quantitative?

- Can we obtain some theory where models can be penalised for having (unjustifiably) specific predictions?
- Yes!
- Also, we can resolve the sure-thing “paradox” without ever having to introduce those other 999,999 models into the hypothesis space!
- May have more general applications

# Bayesian Inference

Before we learn the data, we don't know the parameters *or the data*. So we have some prior knowledge described by

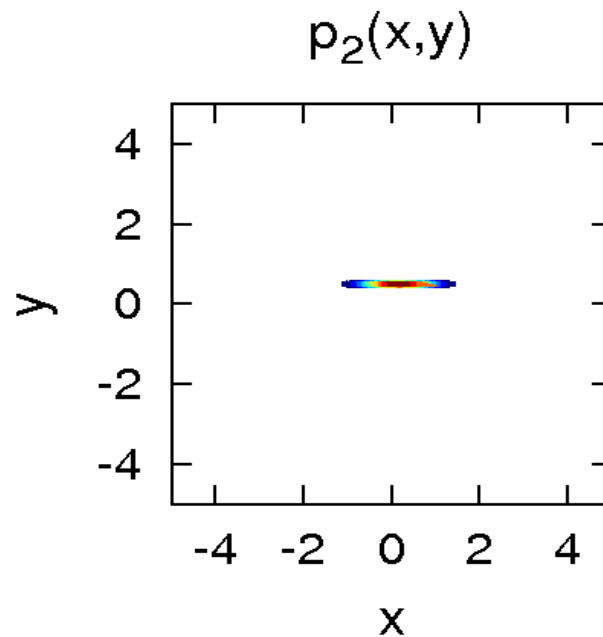
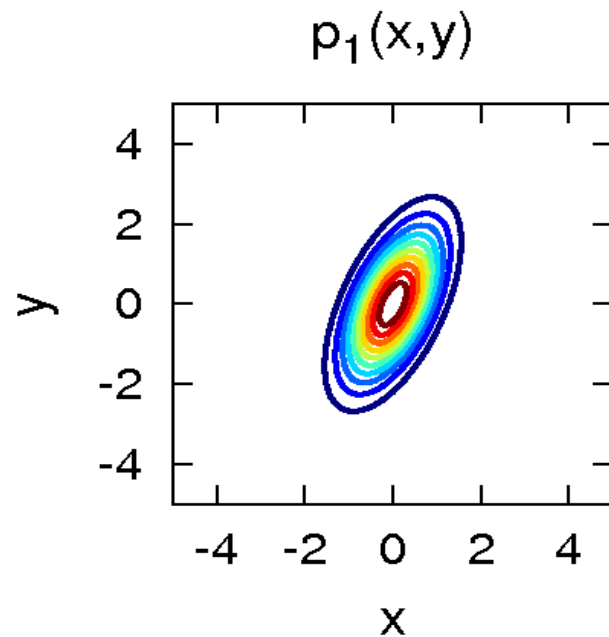
$$p_1(\theta, D) = p_1(\theta)p_1(D|\theta)$$

- Likelihoods describe prior information. They're just as “prior” as the prior!
- The fullest description of our knowledge is a distribution on the product space of possible parameters and data.

# Bayesian Updating

We get specific data  $D_{\text{obs}}$  and update to the posterior.

- Let's look at this in the joint space.



# Why use Bayes' Rule?

- We could just look at the data and write down a posterior.
- We like to go via the “prior, then update” path because at least then we're sure that we got the updating rule right!
- There's another updating rule, maximum relative entropy (MrE), that updates when we learn constraints on allowed probability distributions. See papers by Adom Giffin and Ariel Caticha.

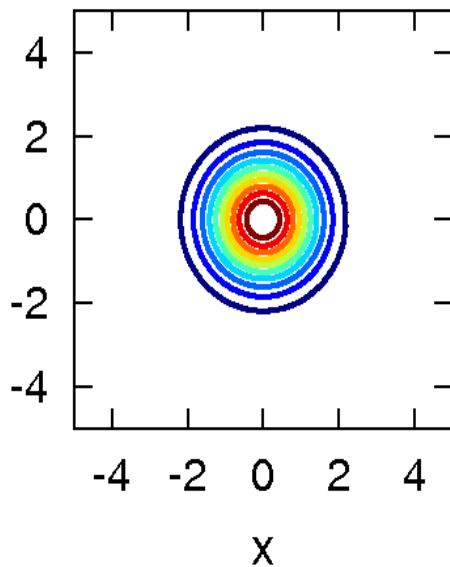
# MrE (huh, yeah), what is it good for?

- Absolutely nothing (Neal)
- Good for updating given constraints that refer directly to our probabilities, but this pretty much never happens in real problems (Skilling?)
- Can summarise relevant external data by the effect on our probabilities.
- e. g. Sampling distributions  $p(D|\theta)$  come from knowledge of the experiment, **possibly preliminary tests of the apparatus with known inputs**. Hardly anyone ever does Bayes on this.

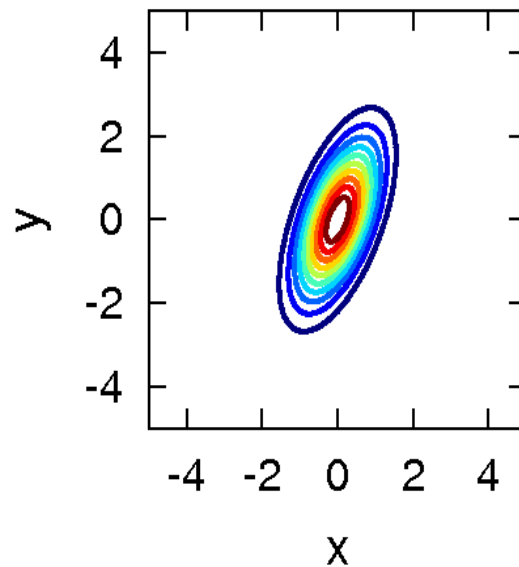
# Entropic Prior

- Knowledge at *three* stages. Update once when given sampling distributions, then again given data.
- In the update from  $p_1$  to  $p_2$ , MrE == Bayes

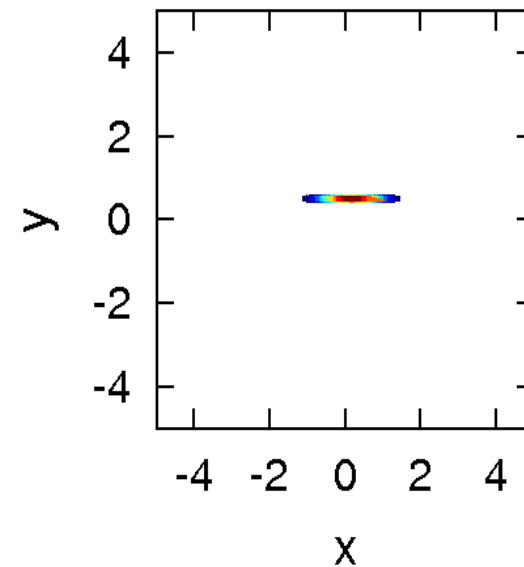
$p_0(x,y)$



$p_1(x,y)$



$p_2(x,y)$



# Entropic Prior

$$p_0(\theta, D) = p_0(\theta)p_0(D)$$

$$S = - \int \int p_1(\theta, D) \log \frac{p_1(\theta, D)}{p_0(\theta, D)} d\theta dD$$

$$\implies p_1(\theta) \propto p_0(\theta) e^{S(D|\theta||p_0(D))}$$

- A “predictiveness” factor appears. This supports models/parameter values whose predictions agree with  $p_0(D)$ .

# Sure Thing Again

- Presumably we would have
- $p_0(H_1) = p_0(H_2) = 1/2$
- $p_0(D) = 10^{-6}$  for all  $D$ .
- Then we learn the two hypotheses have the following sampling distributions
- $p_1(D|H_1) = 10^{-6}$  for all  $D$
- $p_1(D|H_2) = 1$  for  $D = 263878$ , 0 otherwise.
- Update using MrE

# Entropic Prior

- Entropies of the sampling distributions wrt  $p_0(D)$ :

$$S(D|H_1) = - \sum_{i=1}^{10^6} 10^{-6} \log \frac{10^{-6}}{10^{-6}} = 0$$

$$S(D|H_2) = -1 \log \frac{1}{10^{-6}} = \log(10^{-6})$$

Thus, the solution to the lottery problem is:

$$\frac{p(H_2|D)}{p(H_1|D)} = \left(\frac{1}{2}\right) \times \left(\frac{10^{-6}}{1}\right) \times \left(\frac{1}{10^{-6}}\right) = 1$$

# What Just Happened

- Solved the sure thing problem without having to think about 999,999 other hypotheses.
- Much easier generalisation to the case where insufficient reason can't be used.
- Protects against post-hoc proposing of fine tuned models
- More serious applications should be coming soon – sorry, I was too slow for it to be ready for this talk!