Physics 509: Bayesian Priors

Scott Oser
Lecture #6

Outline

Last time: we worked a few examples of Bayesian analysis, and saw that Bayes' theorem provides a mathematical justification for the principle known as "Ockham's razor".

Today:

1) Dependence of Bayesian analysis on prior parameterization
2) Advice on how to choose the right prior
3) "Objective priors"---quantifying our ignorance
4) Maximum entropy priors
5) What to do when you don't know how your data is distributed?

Bayes' Theorem

$$P(H|D,I)=\frac{P(H|I)P(D|H,I)}{P(D|I)}$$

Today we want to examine the science and/or art of how you should choose a prior for a Bayesian analysis.

If this were always easy, everyone would probably be Bayesian.

Prior from a prior analysis

The best solution to any problem is to let someone else solve it for you.

If there exist prior measurements of the quantities you need to estimate, why not use them as *your* prior? (Duh!)

Be careful, of course---if you have reason to believe that the previous measurement is actually a mistake (not just a statistical fluctuation) you wouldn't want to include it.

Even the most complicated statistical analysis does not eliminate the need to apply good scientific judgement and common sense.

Dependence on parameterization

Two theorists set out to predict the mass of a new particle

Carla (writes down theory):

"There should be a new particle whose mass is greater than 0 but less than 1, in appropriate units. I have absolutely no other knowledge about the mass, so I'll assume it has equal chances of having any value between zero and 1---i.e. P(m) = 1."

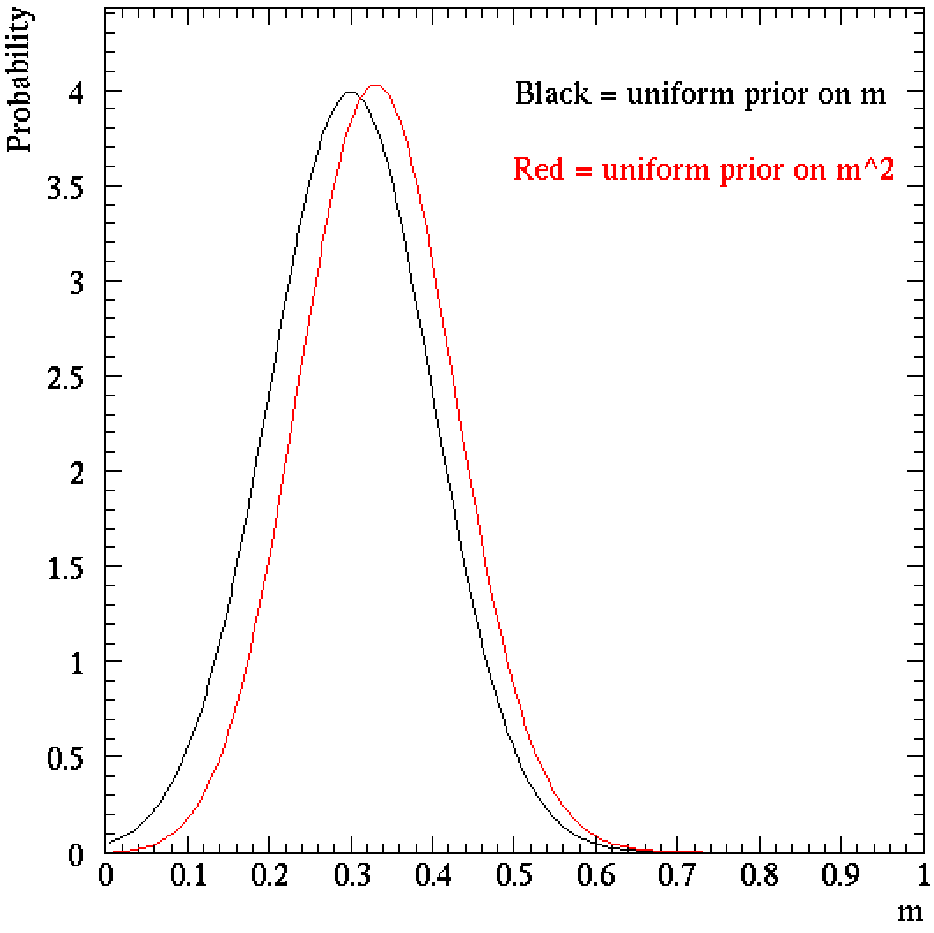
Heidi (writes down the exact same theory):

"There is a new particle described by a single free parameter $y=m^2$ in the Klein-Gordon equation. I'm sure that the true value of y must lie between 0 and 1. Since y is the quantity that appears in my theory, and I know nothing else about it, I'll assume a uniform prior on y---i.e. P(y) = 1."

These are two valid statements of ignorance about the same theory, but with different parameterizations.

An experiment reports: m=0.3±0.1

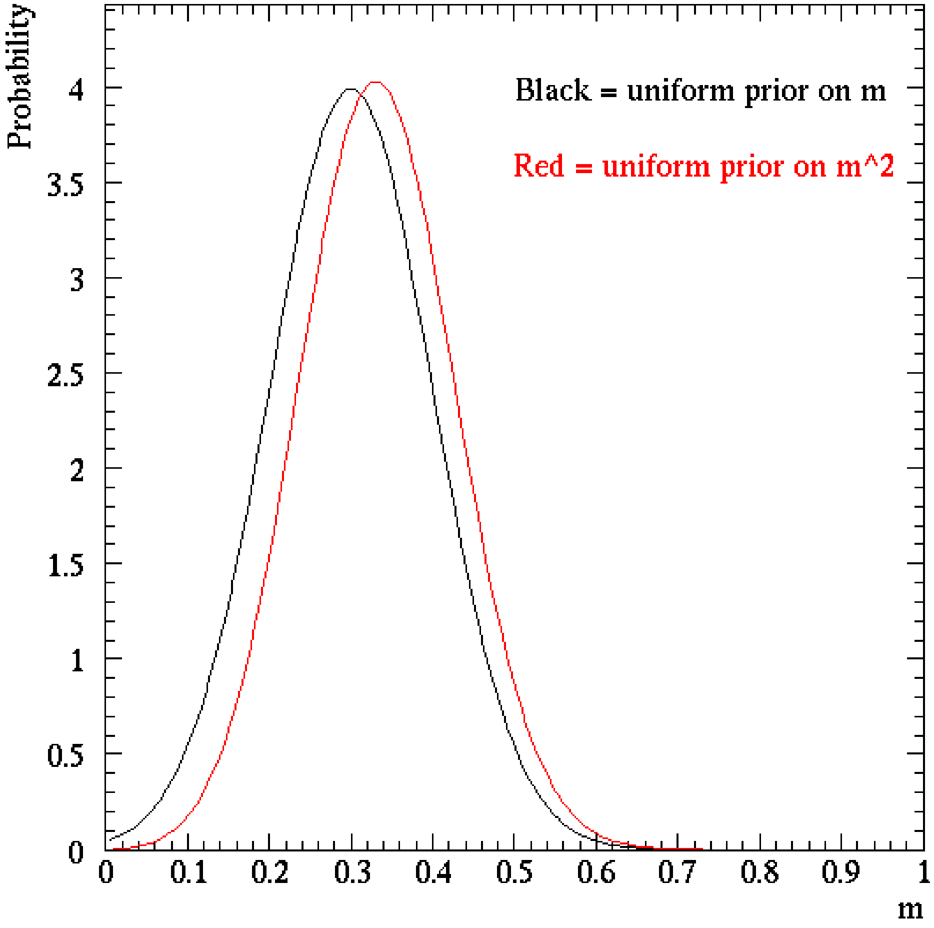
The experimental apparatus naturally measures m, so the experiment reports that (rather than y). Our two theorists incorporate this new knowledge into their theory. Carla calculates a new probability distribution P(m|D,I) for m. Heidi converts the measurement into a statement about the quantity y, and calculates P(y|D,I). They then get together to compare results. Heidi does a change of variables on her PDF so she can directly compare to Carla's result.



The sad truth: choice of parameterization matters

It's quantitatively different to say that all values of m are equally likely versus all values of m^2 are equally likely. The latter will favour larger values of m (if it's 50/50 that m^2 is larger than 0.5, then it's 50/50 than m is larger than 0.707).

Which is right? Statistics alone cannot decide. Only you can, based on physical insight, theoretical biases, etc.



If in doubt, try it both ways.

Principle of Ignorance

In the absence of any reason to distinguish one outcome from another, assign them equal probabilities.

Example: you roll a 6-sided die. You have no reason to believe that the die is loaded. It's intuitive that you should assume that all 6 outcomes are equally likely (p=1/6) until you discover a reason to think otherwise.

Example: a primordial black hole passing through our galaxy hits Earth. We have no reason to believe it's more likely to come from one direction than any other. So we assume that the impact point is uniformly distributed over the Earth's surface.

Parameterization note: this is not the same as assuming that all latitudes are equally likely!

Uniform Prior

Suppose an unknown parameter refers to the location of something (e.g. a peak in a histogram). All positions seem equally likely.

Imagine shifting everything by $x'=x+c$. We demand that $p(X|I)\, dX = P(X'|I)\, dX' = P(X'|I)\, dX$. This is only true for all c if $P(X)$ is a constant.

Really obvious, perhaps ... if you are completely ignorant about the location of something, use a uniform prior for your initial guess of that location.

Note: although a properly normalized uniform prior has a finite range, you can often get away with using a uniform prior from $-\infty$ to $+\infty$ as long as the product of the prior and the likelihood is finite.

Jeffreys Prior

Suppose an unknown parameter measures the size of something, and that we have no good idea how big the thing will be (1mm? 1m? 1km?). We are ignorant about the *scale*. Put another way, our prior should have the same form no matter what units we use to measure the parameter with. If T'=βT, then

$$P(T|I)\,dT = P(T'|I)\,dT' = p(T'|I)\,\beta\,dT$$

$$P(T|I) = \beta\,P(\beta T|I)\,,\text{which is only true for all }\beta\text{ if}$$

$$P(T|I) = \frac{\text{constant}}{T}$$

Properly normalized from T_{min} to T_{max} this is:

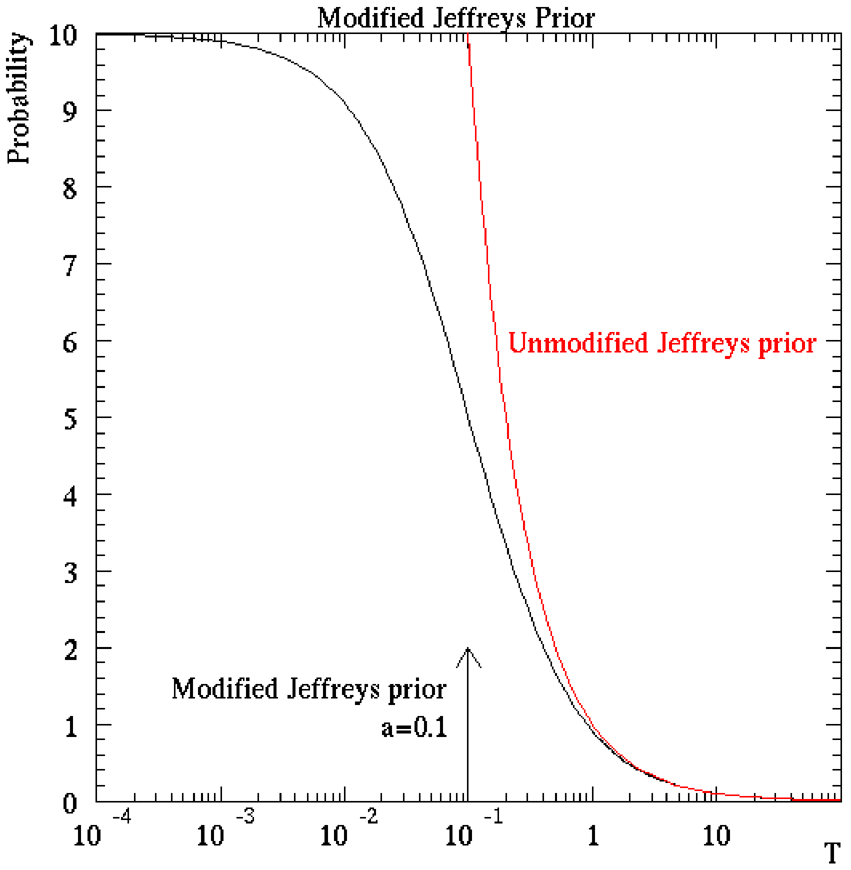
$$P(T|I) = \frac{1}{T\ln(T_{max}/T_{min})}$$

Modified Jeffreys Prior

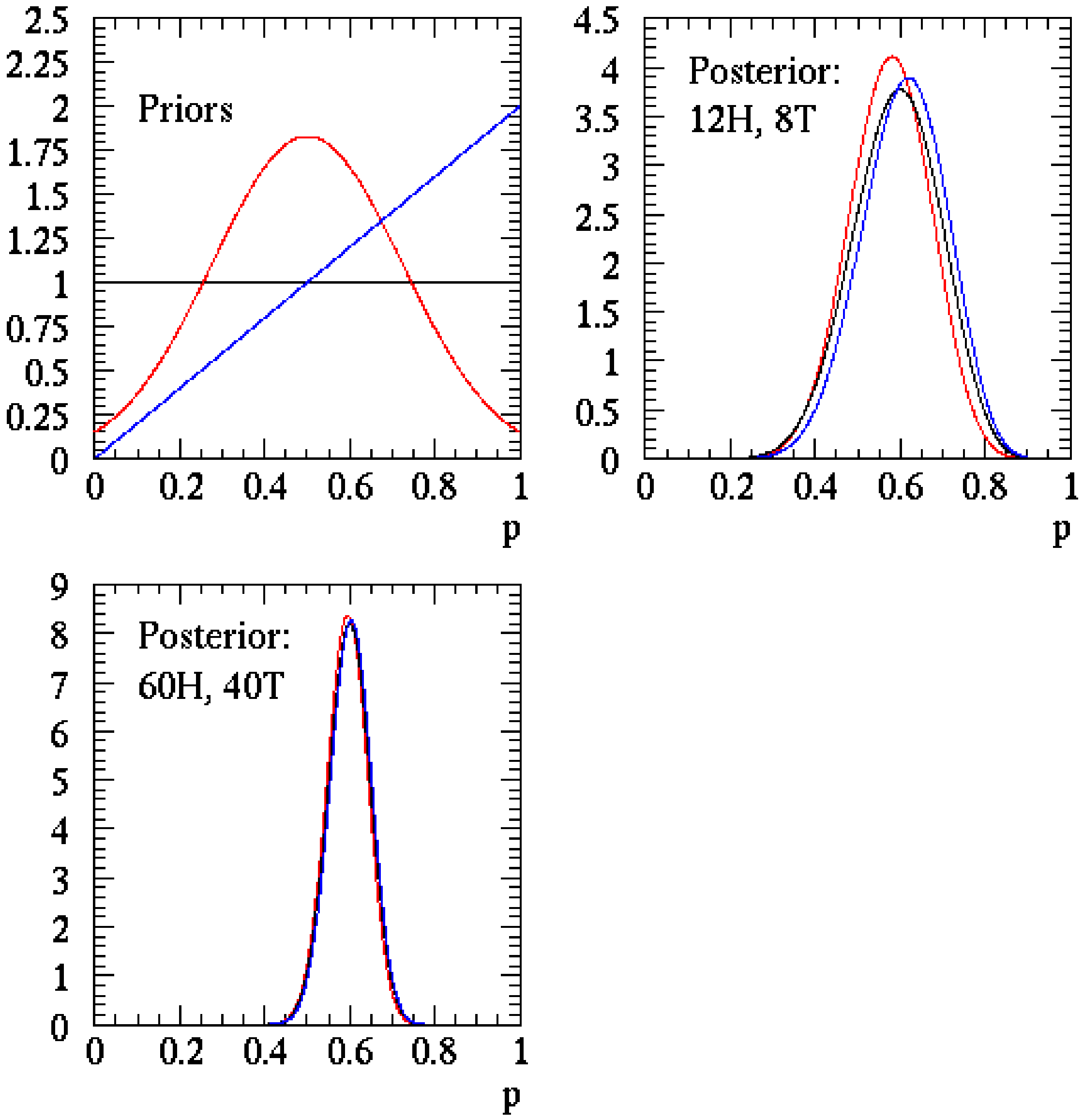
What if your parameter could equal zero? Jeffreys prior is not normalizable---it blows up for T_{min}=0, with probability 1 that T<ε with ε arbitrarily small.

An alternate is the modified Jeffreys prior --- becomes a uniform prior for T < a.

$$P(T|I) = \frac{1}{(T+a)\ln[(a+T_{max})/a]}$$



Given enough data, priors don't matter



The more constraining your data becomes, the less the prior matters.

When posterior distribution is your much narrower than prior, the prior won't vary much over the region of interest. Most priors approximate to flat in this case.

Consider the case of estimating *p* for a binomial distribution after observing 20 or 100 coin flips.

A prior gotcha

Maybe an obvious point ... if your prior ever equals zero at some value, then your posterior distribution must equal zero at that value as well, no matter what your data says.

Be cautious about choosing priors that are identically zero over any range of interest.

“Objective” priors

Much criticism of Bayesian analysis concerns the fact that the result of the analysis depends on the choice of prior, and that the assignment of this prior seems rather subjective.

Is there some objective way of assigning a prior in the case that we know little about its possible distribution?

Shannon's Entropy theorem

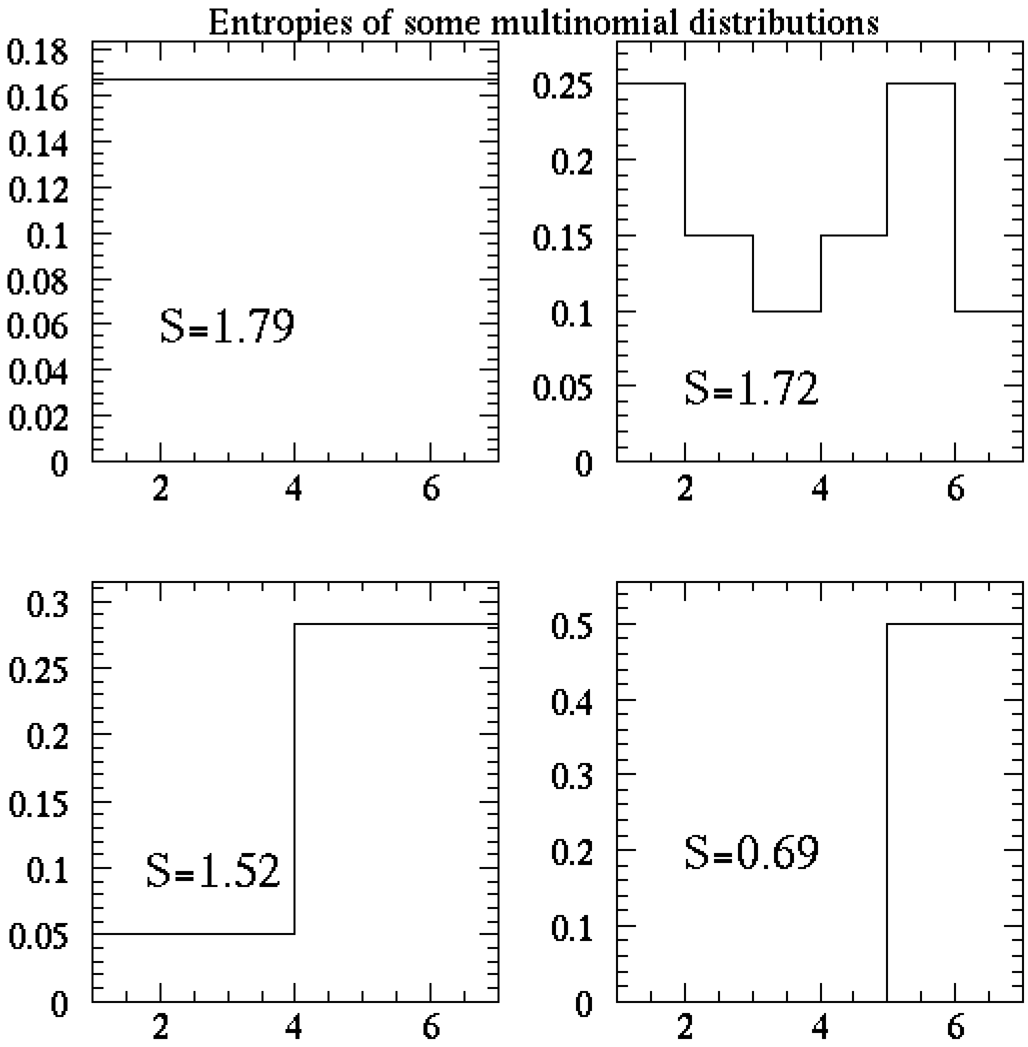
1948: remarkable paper by Charles Shannon generalizes the concept of entropy to a probability distribution:

$$S(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \ln(p_i)$$

Some remarkable properties:

1) The entropy has the same form as the thermodynamic equivalent (modulo Boltzmann's constant).
2) It is a measure of the information content of the distribution. If all the p_i=0 except for one, then S=0---we have a perfect constraint. As uncertainty increases, so does S.
3) The entropy is related to data compression---it is the smallest average number of bits needed to encode a message. (Talk to Colin Gay if you want details.)
4) If S is really a measure of the "information content" of the distribution, and we want to assign a prior that reflects our ignorance of the true value for our parameter, we should assign a prior probability distribution that maximizes S.

Maximum Entropy Principle



The distributions at the left are various probability distributions for the outcomes from a 6-sided die, with the entropy superimposed.

Using the one with the largest entropy as your prior results in the weakest constraint (widest uncertainty) on the posterior PDF.

Maximum Entropy Principle With A Constraint

Often we're not totally ignorant of the prior. For example, perhaps you

- know the mean value of the distribution
- know its variance
- know the average value of some function of the parameter in question

These all are examples of constraints. The maximum entropy prior will then be the probability distribution P(x) that maximizes

$$S=-\sum P_i \ln P_i$$

subject to any constraints that may apply.

Finding the probabilities by a variational method

The mathematical statement of the problem is to find a set of probabilities p_1 ... p_n that maximizes the function

$$S\left(p_1 \ldots p_n\right) = -\sum_{i=1}^{n} p_i \ln p_i$$

If all of the p_i were independent, this would simply imply:

$$dS = \frac{\partial S}{\partial p_1} dp_1 + \frac{\partial S}{\partial p_2} dp_2 + \ldots + \frac{\partial S}{\partial p_n} dp_n = 0$$

Treating the p_i as independent, all of the coefficients must equal zero, and in fact you will wind up concluding that all of the p_i are equal (a uniform prior). This is a mathematical statement of the ignorance principle.

Incorporating constraints with Lagrangian multipliers

Suppose now we impose some constraint on the probability distribution, of the general form C(p_1 ... p_n)=0. Then

$$dC=\frac{\partial C}{\partial p_1}dp_1+\frac{\partial C}{\partial p_2}dp_2+\ldots+\frac{\partial C}{\partial p_n}dp_n=0$$

Therefore $dS-\lambda dC=0$ and so

$$dS-\lambda dC=\left(\frac{\partial S}{\partial p_1}-\lambda\frac{\partial C}{\partial p_1}\right)dp_1+\ldots+\left(\frac{\partial S}{\partial p_n}-\lambda\frac{\partial C}{\partial p_n}\right)dp_n=0$$

We now set the first coefficient equal to zero, giving us an equation for λ, and then we are left with a set of simultaneous equations that can be solved for p_i.

Max Ent prior with only normalization constraint

One constraint always applies: probabilities should sum to 1:

$$C = \sum p_i = 1$$

$$dS - \lambda\, dC = \sum \left(\frac{\partial S}{\partial p_i} - \lambda \frac{\partial C}{\partial p_i} \right) dp_i = \sum \left(-\ln p_i - 1 - \lambda \right) dp_i = 0$$

Allowing the p_i to vary independently and so setting coefficients equal to zero gives:

$$p_i = e^{-(1+\lambda)}$$

We plug back into the constraint equation to determine λ.

Max Ent prior when you know the mean

Suppose we have two constraints---normalization and mean.

$$C = \sum p_i = 1 \qquad \sum y_i p_i = \mu$$

$$d\left[-\sum_{i=1}^{n} p_i \ln p_i - \lambda_0 \left(\sum_{i=1}^{n} p_i - 1\right) - \lambda_1 \left(\sum_{i=1}^{n} y_i p_i - \mu\right)\right] = 0$$

$$\left(-\ln p_i - 1 - \lambda_0 - \lambda_1 y_i\right) = 0$$

$$p_i = e^{-(1+\lambda_0)} e^{-\lambda_1 y_i}$$

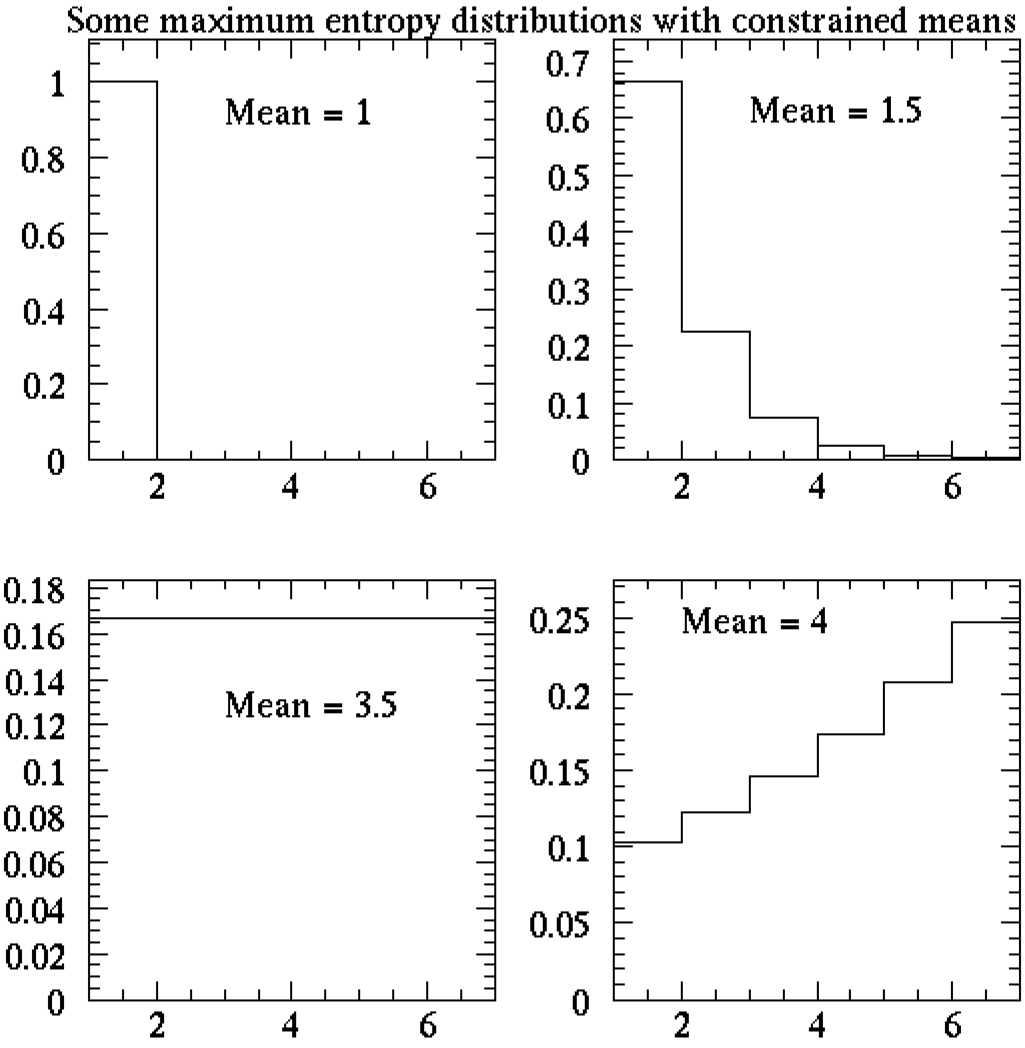
We plug this back into the constraint equations to determine λ_0 and λ_1. The λ_0 factor is a boring normalization term. But the other factor sets the mean value of the distribution.

Max Ent prior: setting the mean

$$p_i = e^{-(1+\lambda_0)} e^{-\lambda_1 y_i}$$

$$\sum_{i=1}^{n} y_i p_i = \mu = \frac{\sum y_i e^{-\lambda_1 y_i}}{\sum e^{-\lambda_1 y_i}}$$

Solve numerically for λ_1.


Some maximum entropy distributions with constrained means

Max Ent prior when you know the variance

What happens if you constrain the variance to equal σ^2? (Let's assume here the mean μ is also known.)

Let's suppose you have prior upper and lower limits on your parameter.

In the limit that the variance is small compared to the range of the parameter:

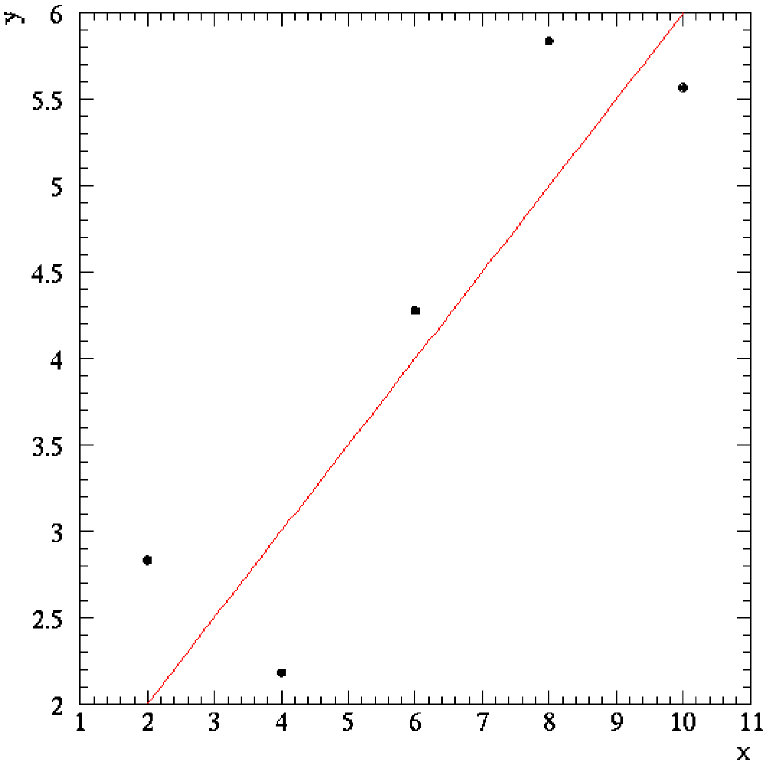
$$\frac{y_{max}-\mu}{\sigma} \gg 1 \quad \text{and} \quad \frac{\mu-y_{min}}{\sigma} \gg 1$$

then the Max Ent distribution with the specified variance is a Gaussian:

$$P(y)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-(y-\mu)^2/2\sigma^2}$$

A Gaussian is the least constraining assumption for the error distribution

A very useful and surprising result follows from this maximum entropy argument. Suppose your data is scattered around your model with an unknown error distribution:



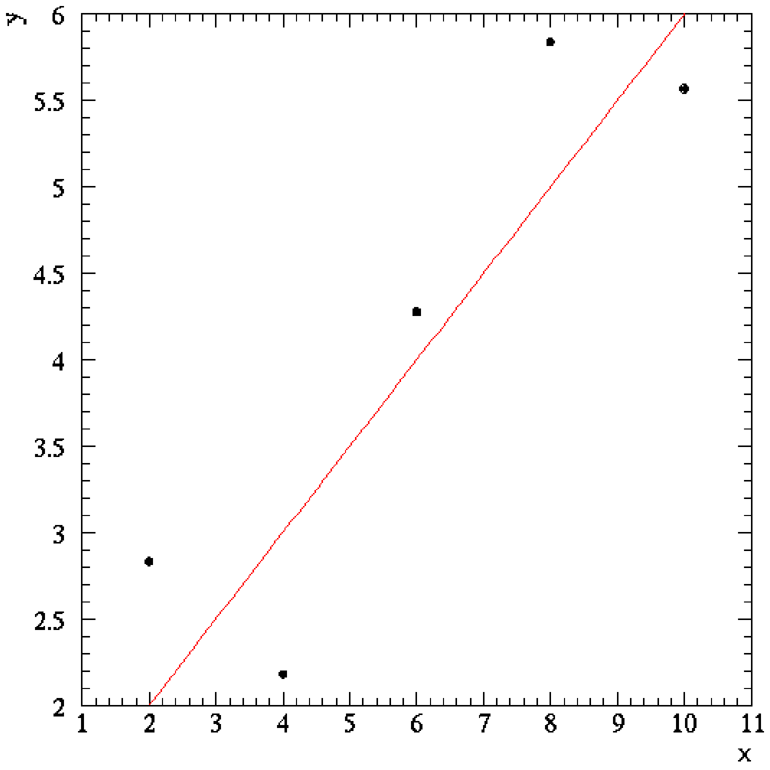
In this example each point is scattered around the model by an error uniformly distributed between -1 and +1.

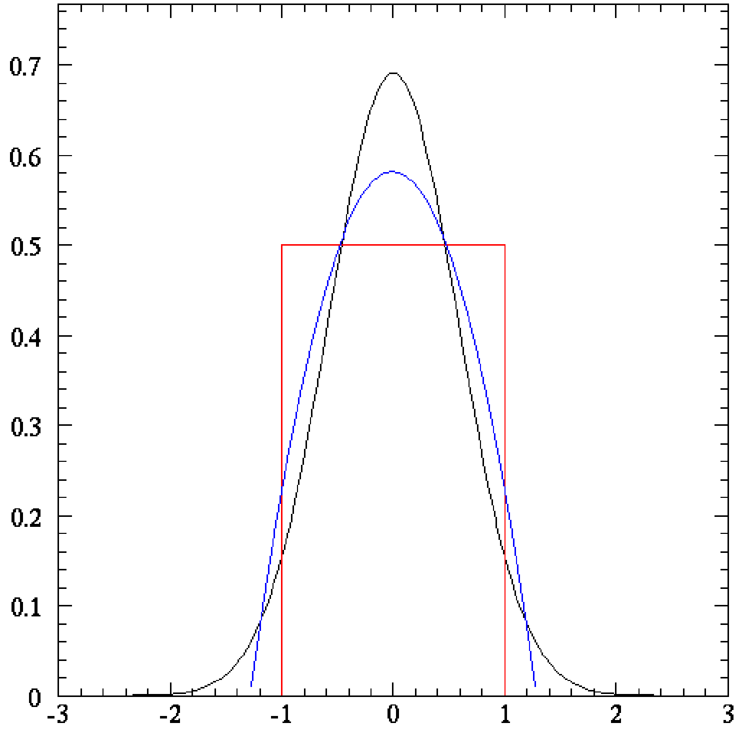
But suppose I don't know how the errors are distributed. What's the most conservative thing I can assume?

A Gaussian error distrib.

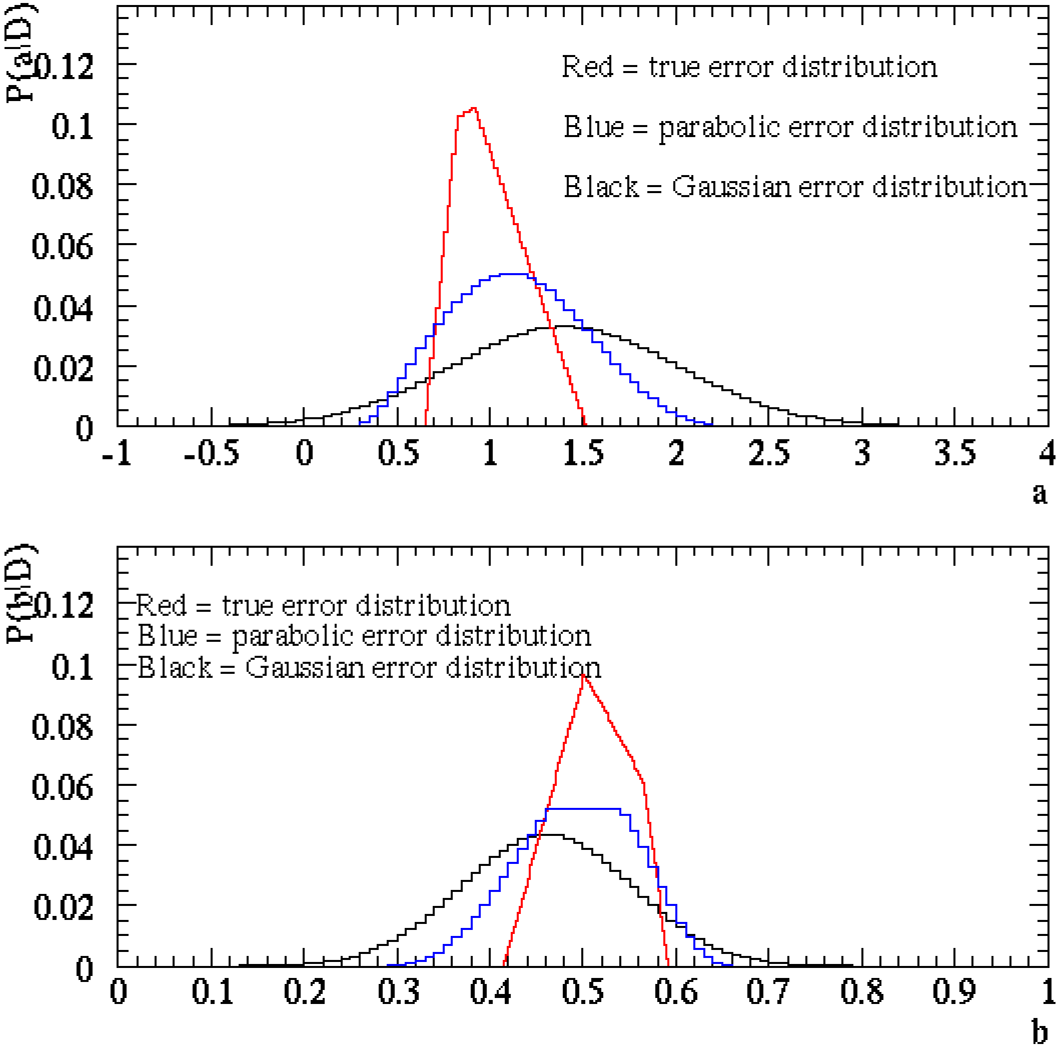
Consider three possible error models

I don't know how the errors are distributed, but I happen to know the RMS of the data around the model by some means. (Maybe Zeus told me.) I consider three possible models for the error: uniform, Gaussian, and parabolic.





Posterior probability distributions for the three error models



These are marginalized PDFs.

Caveat: although in this case the true error distribution gave the tightest parameter constraints, it's perfectly possible for an incorrect assumption about the error distribution to give inappropriately tight constraints!

What if you don't know the RMS?

Imagine that the data is so sparse that you don't already know the scatter of the data around the model.

One possibility is to assume a Gaussian distribution for the errors à la the maximum entropy principle, but to leave σ^2 as a free parameter. Assign it a physically plausible prior (possibly a Jeffreys prior over physically plausible range) and just treat it as a nuisance parameter.

This is more or less like “fitting” for the size of the error.

A very difficult cutting-edge problem ...

Maximum entropy has an important subtlety when dealing with continuous distributions. The continuous case is:

$$S = -\int dx\, p(x) \ln\left(\frac{p(x)}{m(x)}\right)$$

The weird function *m(x)* is really the number density of points in parameter space as you go from the discrete case to the continuous limit.

It's really not obvious what *m(x)* should be. If you know it already, you can use maximum entropy to calculate priors given additional constraints. But if you know absolutely nothing, you can't even define *m(x)*. If you like, *m(x)* is the prior given no constraints at all.

To get beyond this you must use other principles---for example, use transformation symmetries to generate *m(x)*. A common solution is the general Jeffreys prior---choose a prior that is invariant under a parameter transformation.