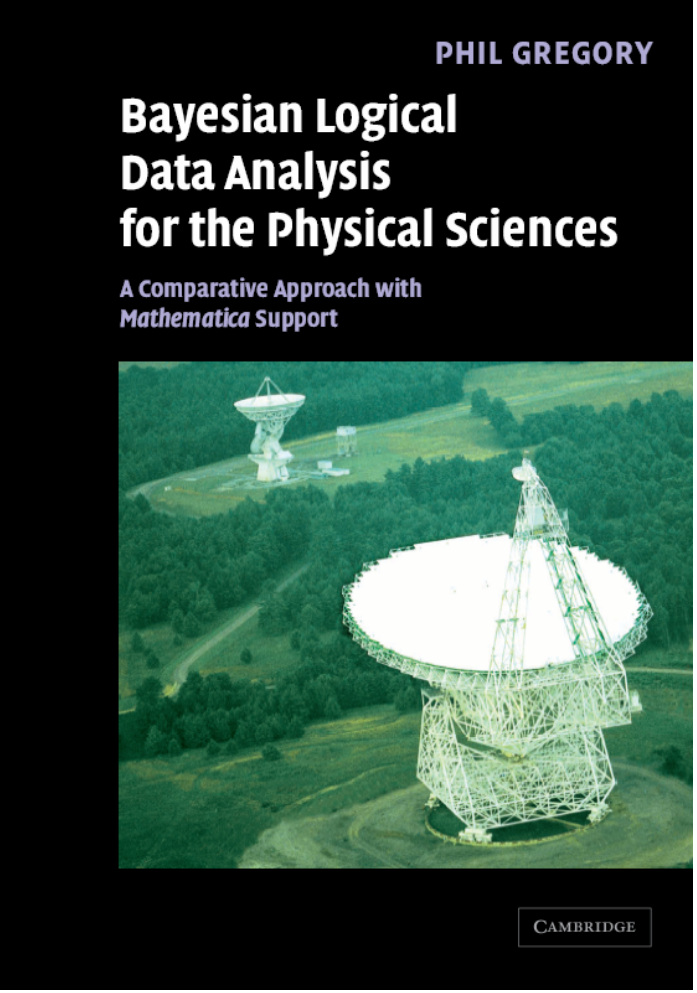
Tutorial on Bayesian Data Analysis

Phil Gregory

University of British Columbia

Nov. 2011



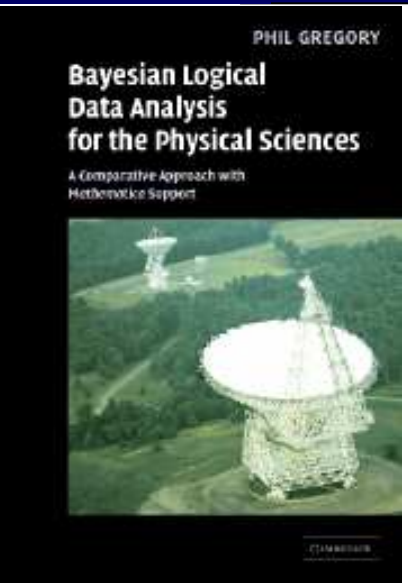
Chapters

1. Role of probability theory in science
2. Probability theory as extended logic
3. The how-to of Bayesian inference
4. Assigning probabilities
5. Frequentist statistical inference
6. What is a statistic?
7. Frequentist hypothesis testing
8. Maximum entropy probabilities
9. Bayesian inference (Gaussian errors)
10. Linear model fitting (Gaussian errors)
11. Nonlinear model fitting
12. Markov chain Monte Carlo
13. Bayesian spectral analysis
14. Bayesian inference (Poisson sampling)

Resources and solutions
This title has free Mathematica based support software available

Introduces statistical inference in the larger context of scientific methods, and includes 55 worked examples and many problem sets.

Hardcover ISBN 9780521841504 | Paperback ISBN:9780521150125)



Resources and solutions

www.cambridge.org/9780521150125

Book Preface (11 Kb)

Errata and Revisions (375 Kb)

Additional book examples with Mathematica 6 tutorial (5.2MB, nb)

Additional book examples with Mathematica 7 tutorial (5.4MB, nb)

Additional book examples with Mathematica 8 tutorial (5.4MB, nb)

Worked solutions for selected problems in Mathematica 7 (7.5MB, nb)

***Mathematica Player*: Free Interactive Player for ...**

Go to Mathematica support material

Outline (Lecture based on Chapter 3 of my book)

Occam's Razor and Model Selection

Occam's razor is principle attributed to the mediaeval philosopher William of Occam . The principle states that one should not make more assumptions than the minimum needed. It underlies all scientific modeling and theory building.

It cautions us to choose from a set of otherwise equivalent models (hypotheses) of a given phenomenon the simplest one.

In any given model, Occam's razor helps us to "shave off" those concepts, variables or constructs that are not really needed to explain the phenomenon.

It was previously thought to be only a qualitative principle. One of the great advantages of Bayesian analysis is that it allows for a quantitative evaluation of Occam's razor.

Occam's Razor and Model Selection

Imagine comparing two models: M_1 with a single parameter, θ, and M_0 with θ fixed at some default value θ_0 (so M_0 has no free parameter).

We will compare the two models by computing the ratio of their posterior probabilities or Odds (O_{10}).

Expand numerator and denominator with Bayes' theorem

$$O_{10} = \frac{p(M_1 \mid D, I)}{p(M_0 \mid D, I)} = \frac{\dfrac{p(M_1 \mid I)\, p(D \mid M_1, I)}{p(D \mid I)}}{\dfrac{p(M_0 \mid I)\, p(D \mid M_0, I)}{p(D \mid I)}} = \frac{p(M_1 \mid I)}{p(M_0 \mid I)} \frac{p(D \mid M_1, I)}{p(D \mid M_0, I)}$$

posterior probability ratio

prior probability ratio

Bayes factor

Occam's Razor and Model Selection

$$\text{Odds} = \frac{p(M_1 \mid D, I)}{p(M_0 \mid D, I)} = \frac{p(M_1 \mid I)}{p(M_0 \mid I)} \frac{p(D \mid M_1, I)}{p(D \mid M_0, I)}$$

Suppose prior Odds = 1

2nd term called Bayes factor B_{10}

$$B_{10} = \frac{p(D \mid M_1, I)}{p(D \mid M_0, I)} = \text{Marginal likelihood ratio}$$

called the Global likelihood ratio in my book

Marginal likelihood for M_1

$$p(D \mid M_1, I) = \int d\theta \; p(\theta \mid M_1, I) \; p(D \mid \theta, M_1, I)$$

In words: the marginal likelihood for a model is the weighted average likelihood for its parameter(s). The weighting function is the prior for the parameter.

To develop our intuition about the Occam penalty we will carry out a back of the envelop calculation for the Bayes factor.

Approximate the prior, $p(\theta|M_1,I)$, by a uniform distribution of width $\Delta\theta$.

Therefore $p(\theta|M_1,I) = 1/\Delta\theta$

$$p(D|M_1,I) = \int d\theta \; p(\theta|M_1,I)\; p(D|\theta,M_1,I)$$

$$= \frac{1}{\Delta\theta}\int d\theta \;\; p(D|\theta,M_1,I)$$

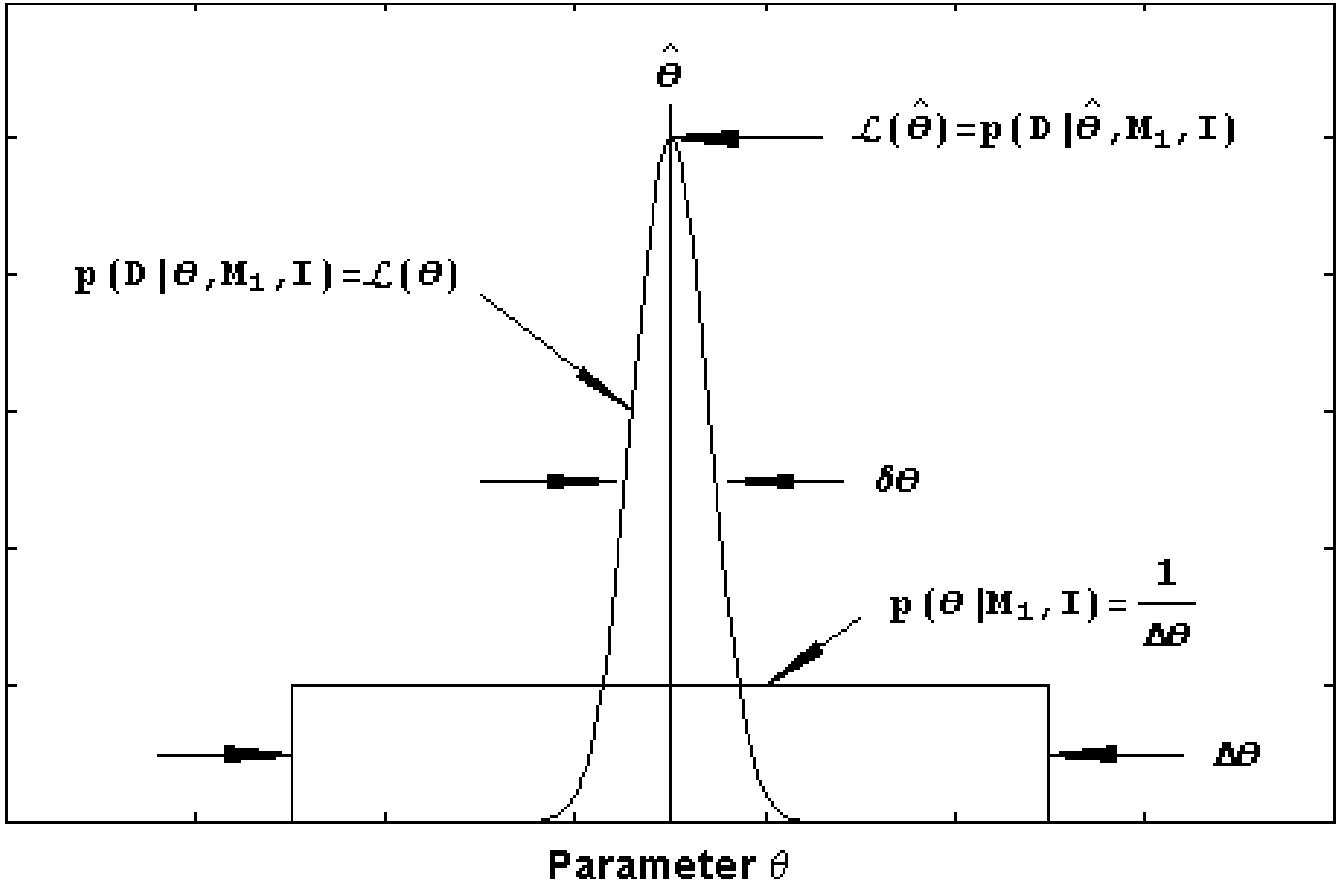
Often the data provide us with more information about parameters than we had without the data, so that the likelihood function, $p(D|\theta,M_1,I)$, will be much more "peaked" than the prior, $p(\theta|M_1,I)$.

Approximate the likelihood function, $p(D|\theta,M_1,I)$, by a Gaussian distribution of a characteristic width $\delta\theta$.

$$p(D|M_1,I) = p(D|\hat{\theta},M_1,I) \;\; \frac{\delta\theta}{\Delta\theta}$$

Maximum value of the likelihood

Occam penalty continued



Since model M_0 has no free parameters the marginal likelihood is also the maximum likelihood, and there is no Occam factor.

$$p(D|M_0, I) = p(D|\theta_0, M_1, I)$$

Now pull the relevant equations together

$$B_{10} = \frac{p(D|M_1, I)}{p(D|M_0, I)} = \frac{p(D|\hat{\theta}, M_1, I)}{p(D|\theta_0, M_0, I)} \times \frac{\delta\theta}{\Delta\theta}$$

Occam penalty continued

Occam factor Ω_θ

$$B_{10} = \frac{p(D \mid \hat{\theta}, M_1, I)}{p(D \mid \theta_0, M_0, I)} \frac{\delta\theta}{\Delta\theta}$$

The maximum likelihood ratio in the first factor can never favor the simpler model because M_1 contains it as a special case.

However, since the posterior width, $\delta\theta$, is narrower than the prior width, $\Delta\theta$, the Occam factor penalizes the complicated model for any "wasted" parameter space that gets ruled out by the data.

The Bayes factor will thus favor the more complicated model only if the likelihood ratio is large enough to overcome this Occam factor.

Suppose M_1 had two parameters θ and φ. Then the Bayes factor would have two Occam factors

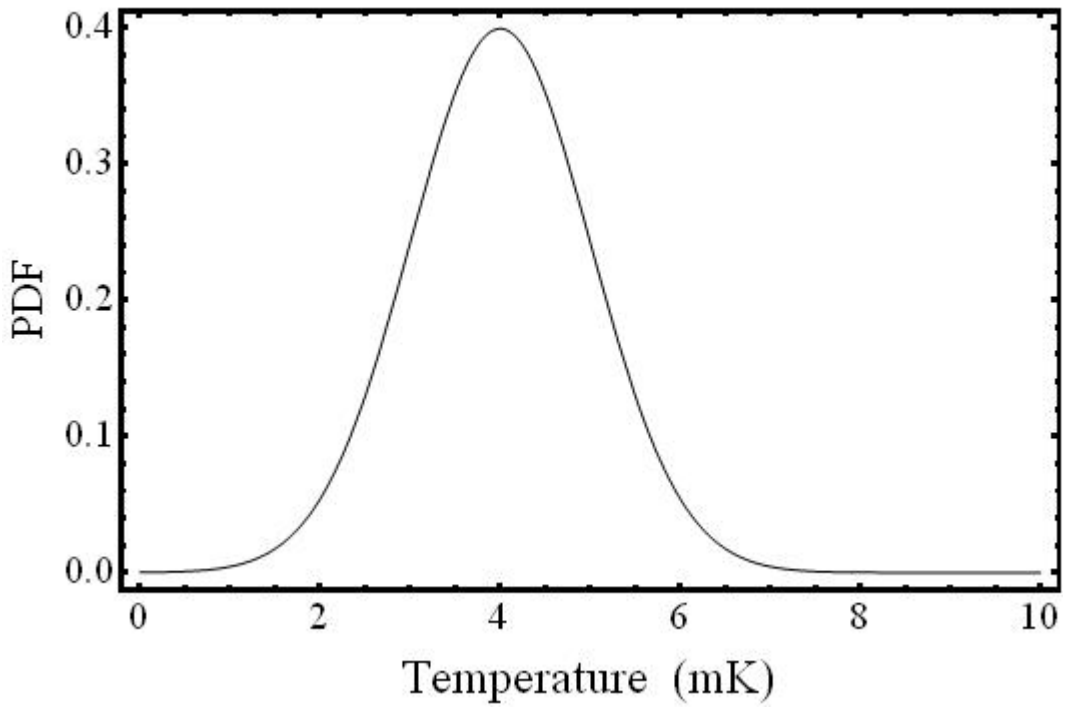
$$B_{10} = \frac{p(D \mid \hat{\theta}, M_1, I)}{p(D \mid \theta_0, M_0, I)} \frac{\delta\theta}{\Delta\theta} \frac{\delta\varphi}{\Delta\varphi}$$

$$= \text{max likelihood ratio} \times \Omega_\theta \times \Omega_\varphi$$

Occam,s razor continued

Note: parameter estimation is like model selection only we are comparing models with the same complexity so the Occam factors cancel out.

Warning: do not try and use a parameter estimation analysis to do model selection.



Simple Spectral Line Problem

Background (prior) information:

Two competing grand unification theories have been proposed, each championed by a Nobel prize winner in physics. We want to compute the relative probability of the truth of each theory based on our prior information and some new data.

Theory 1 is unique in that it predicts the existence of a new short-lived baryon which is expected to form a short-lived atom and give rise to a spectral line at an accurately calculable radio wavelength.

Unfortunately, it is not feasible to detect the line in the laboratory. The only possibility of obtaining a sufficient column density of the short-lived atom is in interstellar space.

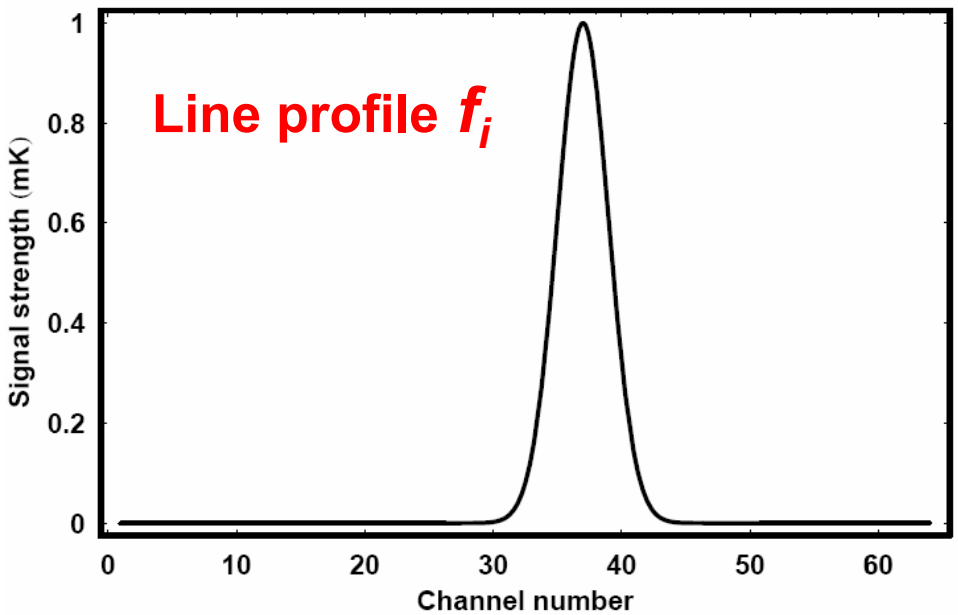
Prior estimates of the line strength expected from the Orion nebula according to theory 1 range from 0.1 to 100 mK.

Simple Spectral Line Problem

The predicted line shape has the form

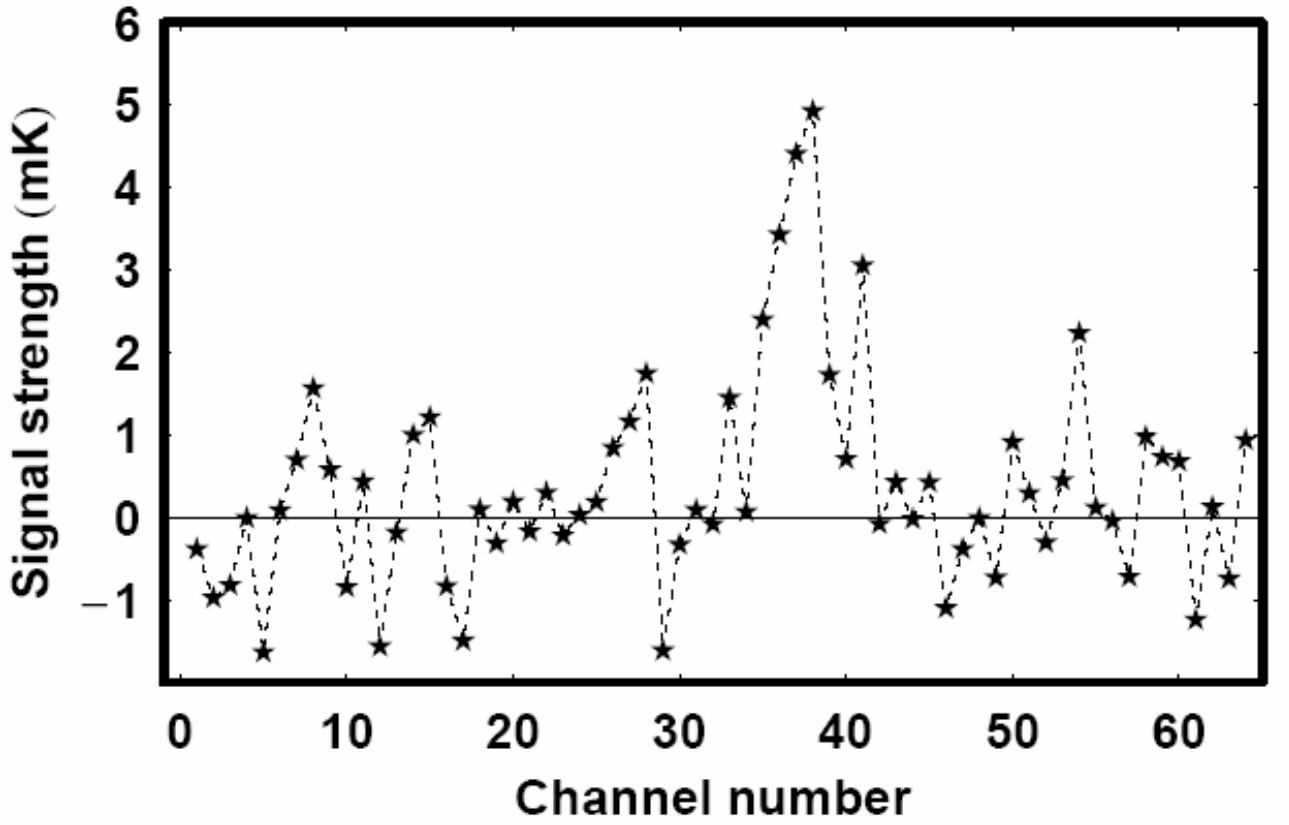
$$T \exp\left\{\frac{-(\nu_i - \nu_o)^2}{8}\right\} \quad \text{(abbreviated by } T f_i\text{)},$$

where the signal strength is measured in temperature units of mK and *T* is the amplitude of the line. The frequency, ν_i , is in units of the spectrometer channel number and the line center frequency $\nu_0 = 37$.



Data

To test this prediction, a new spectrometer was mounted on the James Clerk Maxwell telescope on Mauna Kea and the spectrum shown below was obtained. The spectrometer has 64 frequency channels.



All channels have Gaussian noise characterized by $\sigma = 1$ mK. The noise in separate channels is independent. The line center frequency $\nu_0 = 37$.

Questions of interest

Based on our current state of information, which includes just the above prior information and the measured spectrum,

1) what do we conclude about the relative probabilities of the two competing theories

and

2) what is the posterior PDF for the model parameters?

Hypothesis space of interest for model selection part:

M_0 ≡ "Model 0, no line exists"
M_1 ≡ "Model 1, line exists"

M_1 has 1 unknown parameters, the line temperature *T*.

M_0 has no unknown parameters.

Model selection

To answer the model selection question, we compute the odds ratio (O_{10}) of model M_1 to model M_0 .

Expand numerator and denominator with Bayes' theorem

$$O_{10} = \frac{p(M_1 \mid D, I)}{p(M_0 \mid D, I)} = \frac{\dfrac{p(M_1 \mid I)\, p(D \mid M_1, I)}{p(D \mid I)}}{\dfrac{p(M_0 \mid I)\, p(D \mid M_0, I)}{p(D \mid I)}} = \frac{p(M_1 \mid I)}{p(M_0 \mid I)} \frac{p(D \mid M_1, I)}{p(D \mid M_0, I)}$$

posterior probability ratio

prior probability ratio

Bayes factor

$p(D|M_1, I)$, the called the marginal (or global) likelihood of M_1 .

$$p(D \mid M_1, I) = \int_T p(D, T \mid M_1, I)\, dT$$

Expanded with product rule

$$= \int_T p(T \mid M_1, I)\; p(D \mid M_1, T, I)\, dT$$

The marginal likelihood of a model is equal to the weighted average likelihood for its parameters.

Choice of prior $p(T|M_1, I)$

Investigate two common choices

1. Uniform prior

$$p(T|M_1, I) = \frac{1}{\Delta T}$$

$$\text{where } \Delta T = T_{\max} - T_{\min}$$

There is a problem with this prior if the range of *T* is large. In the current example T_{min} = 0.1 and T_{max} = 100. Compare the probability that *T* lies in the upper decade of the prior range (10 to 100 mK) to the lowest decade (0.1 to 1 mK).

$$\frac{\int_{10}^{100} p(T|M_1, I)dT}{\int_{0.1}^{1} p(T|M_1, I)dT} = 100$$

Usually, expressing great uncertainty in some quantity corresponds more closely to a statement of scale invariance or equal probability per decade. The Jeffreys prior, discussed next, has this scale invariant property.

2. Jeffreys scale invariant prior

$$p(T \mid M_1, I)\,dT = \frac{dT}{T \times \ln(T_{max}/T_{min})}$$

or equivalently $$p(\ln T \mid M_1, I)\,d\ln T = \frac{d\ln T}{\ln(T_{max}/T_{min})}$$

$$\int_{0.1}^{1} p(T \mid M_1, I)\,dT = \int_{10}^{100} p(T \mid M_1, I)\,dT$$

What if the lower bound on T includes zero? Another alternative Is a modified Jeffreys prior of the form.

$$p(T \mid M_1, I) = \frac{1}{T + T_0} \frac{1}{\ln\left(\frac{T_0 + T_{max}}{T_0}\right)}$$

This prior behaves like a uniform prior for $T < T_0$ and a Jeffreys prior for $T > T_0$. Typically set T_0 = noise level.

Calculation of $p(D|M_1, T, I)$

Let d_i represent the measured data value for the i^{th} channel of the spectrometer. According to model M_1,

$$d_i = T f_i + e_i \quad \text{and} \quad f_i = \exp\left(\frac{-(\nu_i - \nu_0)^2}{2\sigma_L^2}\right),$$

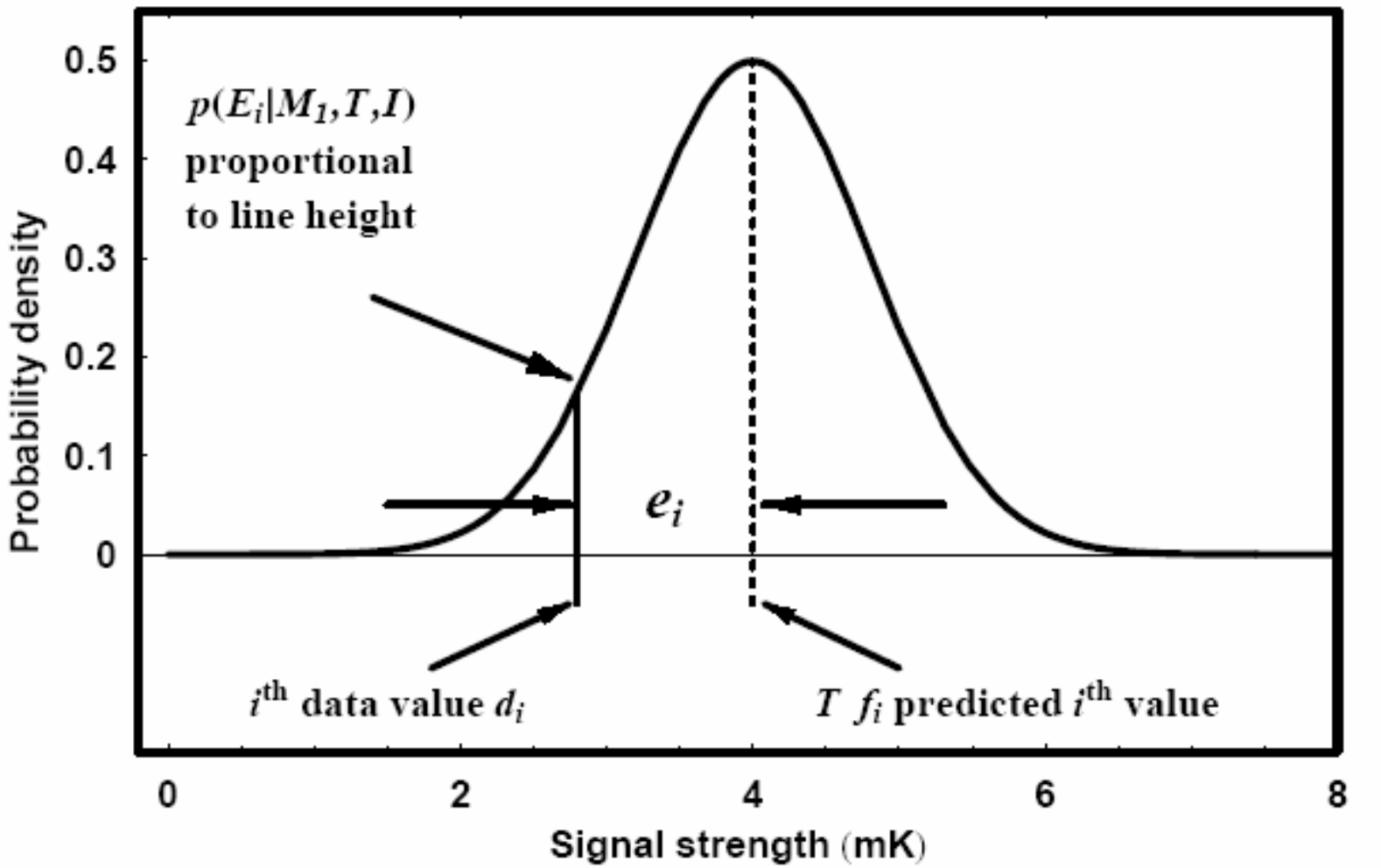
and e_i represents the error component in the measurement. Our prior information indicates that e_i has a Gaussian distribution with a $\sigma = 1$ mK.

Assuming M_1 is true, then if it were not for the error e_i d_i would equal the model prediction $T f_i$.

Let $E_i \equiv$ "a proposition asserting that the i^{th} error value is in the range e_i to $e_i + de_i$." If all the E_i are independent then

$$\begin{aligned} p(D|M_1, T, I) &= p(D_1, D_2, ..., D_N|M_1, T, I) \\ &= p(E_1, E_2, ..., E_N|M_1, T, I) \\ &= p(E_1|M_1, T, I)p(E_2|M_1, T, I)...p(E_N|M_1, T, I) \\ &= \prod_{i=1}^{N} p(E_i|M_1, T, I) \end{aligned}$$

Calculation of $p(D|M_1, T, I)$



Probability of getting a data value d_i a distance e_i away from the predicted value is proportional to the height of the Gaussian error curve at that location.

Calculation of $p(D|M_1, T, I)$

From the prior information, we can write

$$p(E_i|M_1, T, I) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{e_i^2}{2\sigma^2}\right\}$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(d_i - T f_i)^2}{2\sigma^2}\right\}$$

Our final likelihood is given by

$$p(D \mid M_1, I) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} Exp\left[-\frac{(d_i - T\ f_i)^2}{2\sigma^2}\right]$$

$$= (2\pi)^{-N/2} \sigma^{-N} Exp\left[-0.5 \sum_{i=1}^{N} \frac{(d_i - T\ f_i)^2}{\sigma^2}\right]$$

The familiar χ^2 statistic used in least-squares

Maximum and Marginal likelihoods

Our final likelihood is given by

$$p(D \mid M_1, I) = (2\pi)^{-N/2}\, \sigma^{-N}\, Exp\left[-\sum_{i=1}^{N} \frac{(d_i - T\ f_i)^2}{2\sigma^2}\right]$$

For the given data the maximum value of the likelihood = 3.80×10^{-37}

To compute the odds O_{10} need the marginal likelihood of M_1 .

$$p(D \mid M_1, I) = \int_T p(T \mid M_1, I)\ p(D \mid M_1, T, I)\, dT$$

A uniform prior for *T* yields $p(D|M_1,I) = 5.06\times10^{-39}$

A Jeffreys prior for *T* yields $p(D|M_1,I) = 1.74\times10^{-38}$

Calculation of $p(D|M_0,I)$

Model M_0 assumes the spectrum is consistent with noise and has no free parameters so we can write

$$d_i = 0 + e_i$$

$$p(D \mid M_0, I) = (2\pi)^{-\frac{N}{2}} \sigma^{-N} \, Exp\left[-\sum_{i=1}^{N} \frac{(d_i - 0)^2}{2\sigma^2} \right] = 3.26\times10^{-51}$$

Model selection results

Bayes factor, uniform prior = 1.6×10^{12}

Bayes factor, Jeffreys prior = 5.3×10^{12}

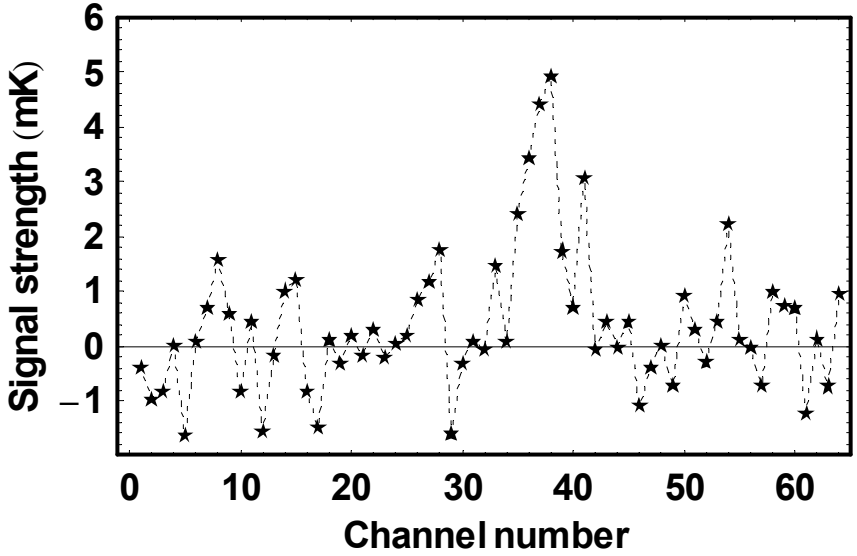
The factor of 10^{12} is so large that we are not terribly interested in whether the factor in front is 1.6 or 5.3. Thus the choice of prior is of little consequence when the evidence provided by the data for the existence is as strong as this.

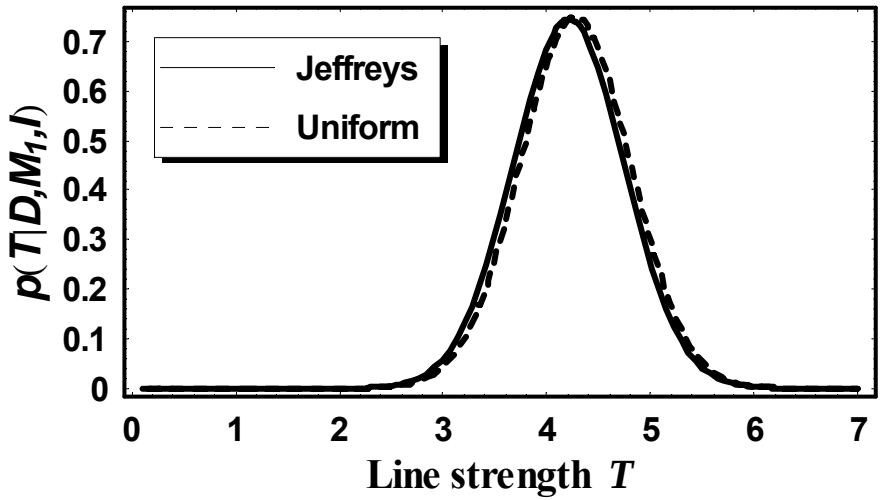
Parameter Estimation Problem

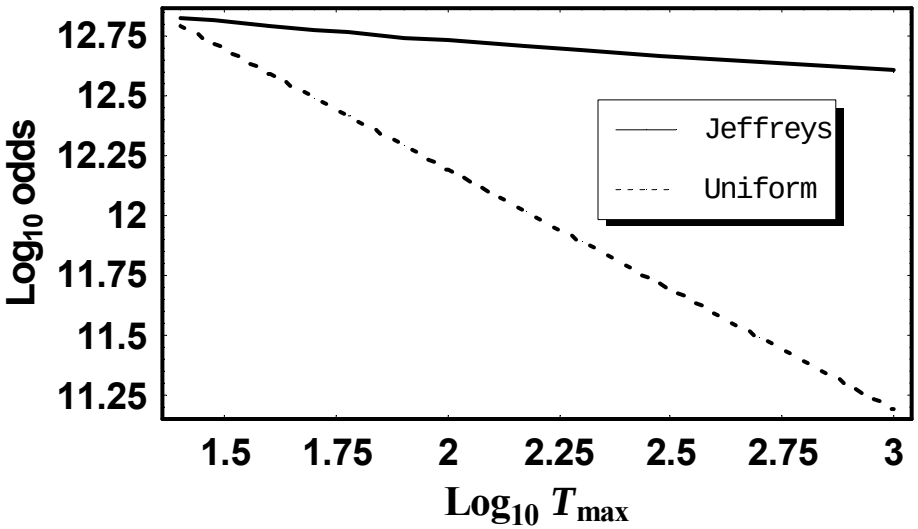
Now that we have solved the model selection problem leading to a significant preference for M_1, we would now like to compute $p(T|D,M_1,I)$, the posterior PDF for the signal strength.

Again, start with Bayes' theorem

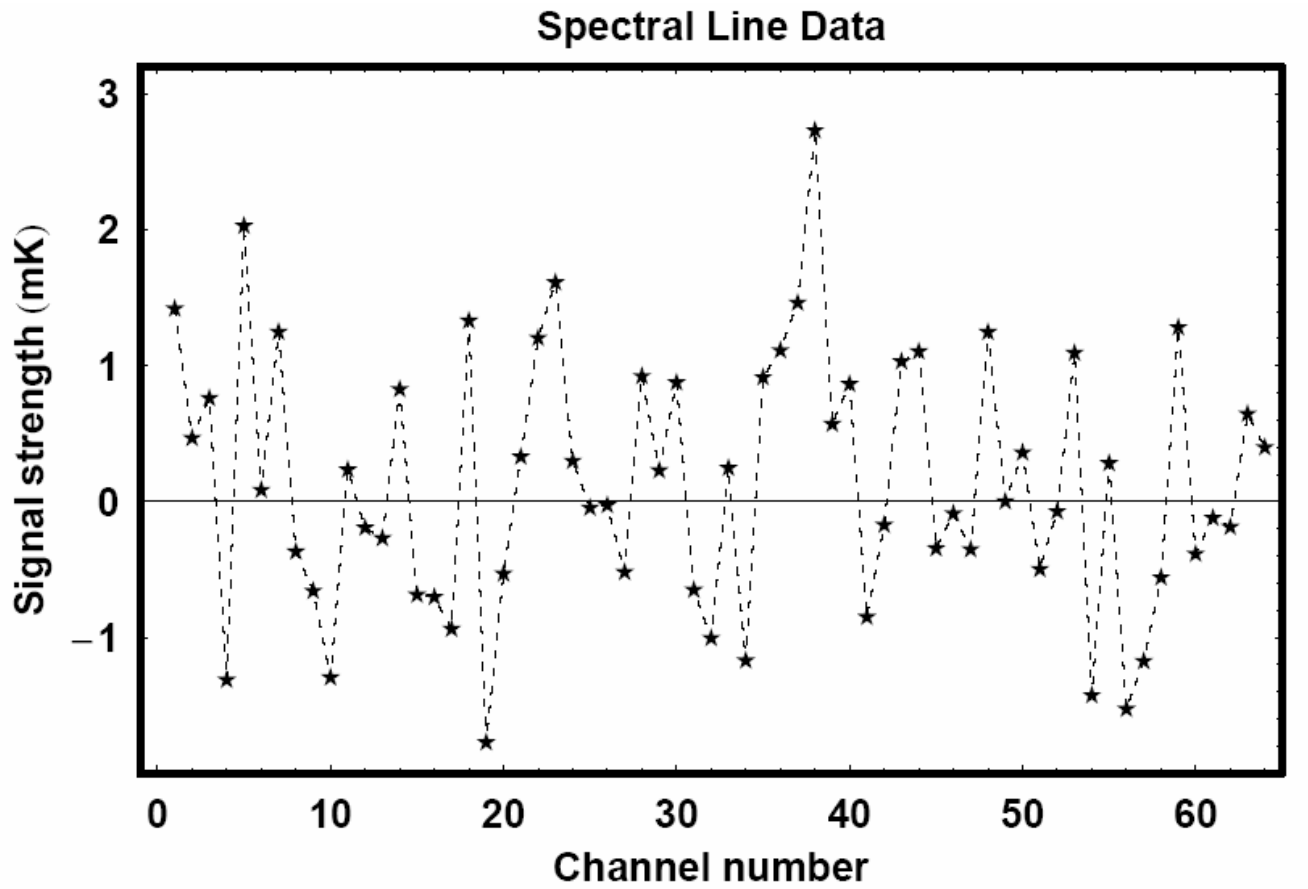
$$p(T|D,M_1,I) = \frac{p(T|M_1,I)p(D|M_1,T,I)}{p(D|M_1,I)}$$
$$\propto p(T|M_1,I)p(D|M_1,T,I)$$


Signal strength (mK)
6
5
4
3
2
1
0
−1
0
10
20
30
40
50
60
Channel number


p(T|D,M1,I)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Jeffreys
Uniform
0
1
2
3
4
5
6
7
Line strength T


Log10 odds
12.75
12.5
12.25
12
11.75
11.5
11.25
Jeffreys
Uniform
1.5
1.75
2
2.25
2.5
2.75
3
Log10 Tmax

How do our conclusions change when evidence for the line in the data is weaker?



All channels have IID Gaussian noise characterized by $\sigma = 1$ mK.
The predicted line center frequency $\nu_0 = 37$.

Weak line model selection results

For model M_0 , $p(D|M_0,I)$ = 1.13×10^{-38}

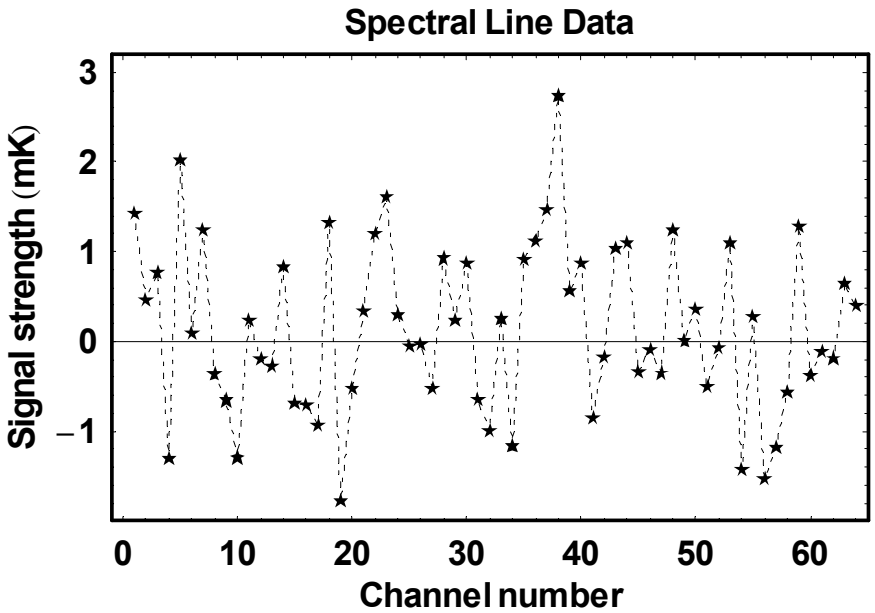
A uniform prior for T yields $p(D|M_1,I)$ = 1.13×10^{-38}

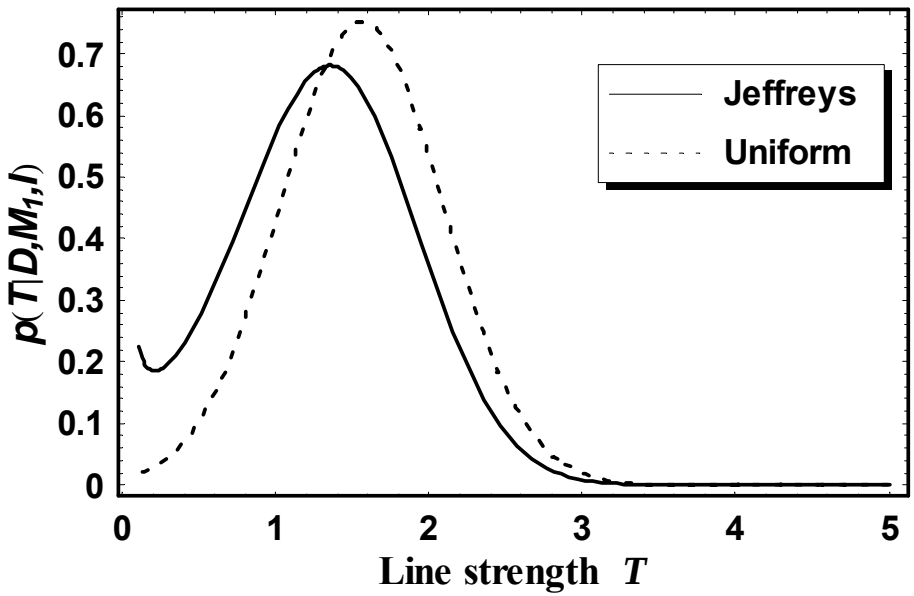
A Jeffreys prior for T yields $p(D|M_1,I)$ = 1.24×10^{-37}

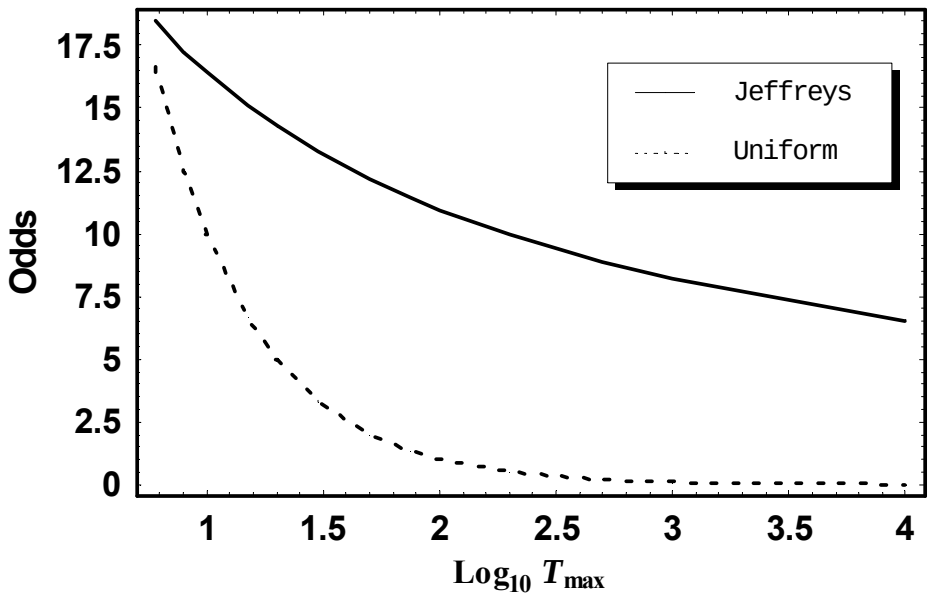
Bayes factor, uniform prior = 1.0

Bayes factor, Jeffreys prior = 11.

As expected, when the evidence provided by the data is much weaker, our conclusions can be strongly influenced by the choice of prior and it is a good idea to examine the sensitivity of the results by employing more than one prior.







What if we were uncertain about the line center frequency?

Suppose our prior information only restricted the center frequency to the first 44 channels of the spectrometer.

In this case ν_0 becomes a nuisance parameter that we can marginalize.

$$p(D \mid M_1, I) = \int_{\nu_0}\int_{T} p(D, T, \nu_0 \mid M_1, I)\, dT\, d\nu_0$$

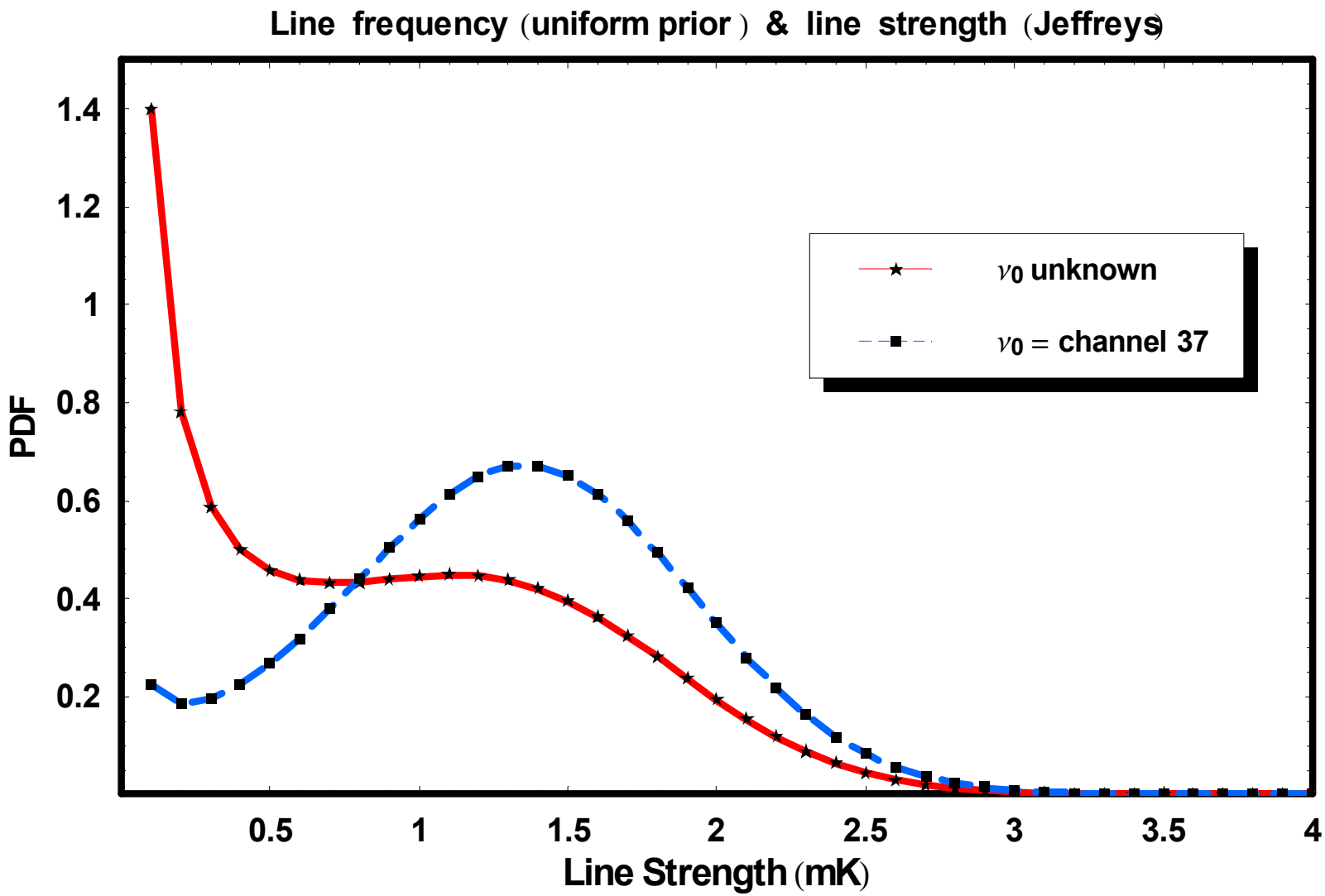
Assumes independent priors

$$= \int_{\nu_0}\int_{T} p(T \mid M_1, I)\; p(\nu_0 \mid M_1, I)\; p(D \mid M_1, T, \nu_0, I)\, dT\, d\nu_0$$

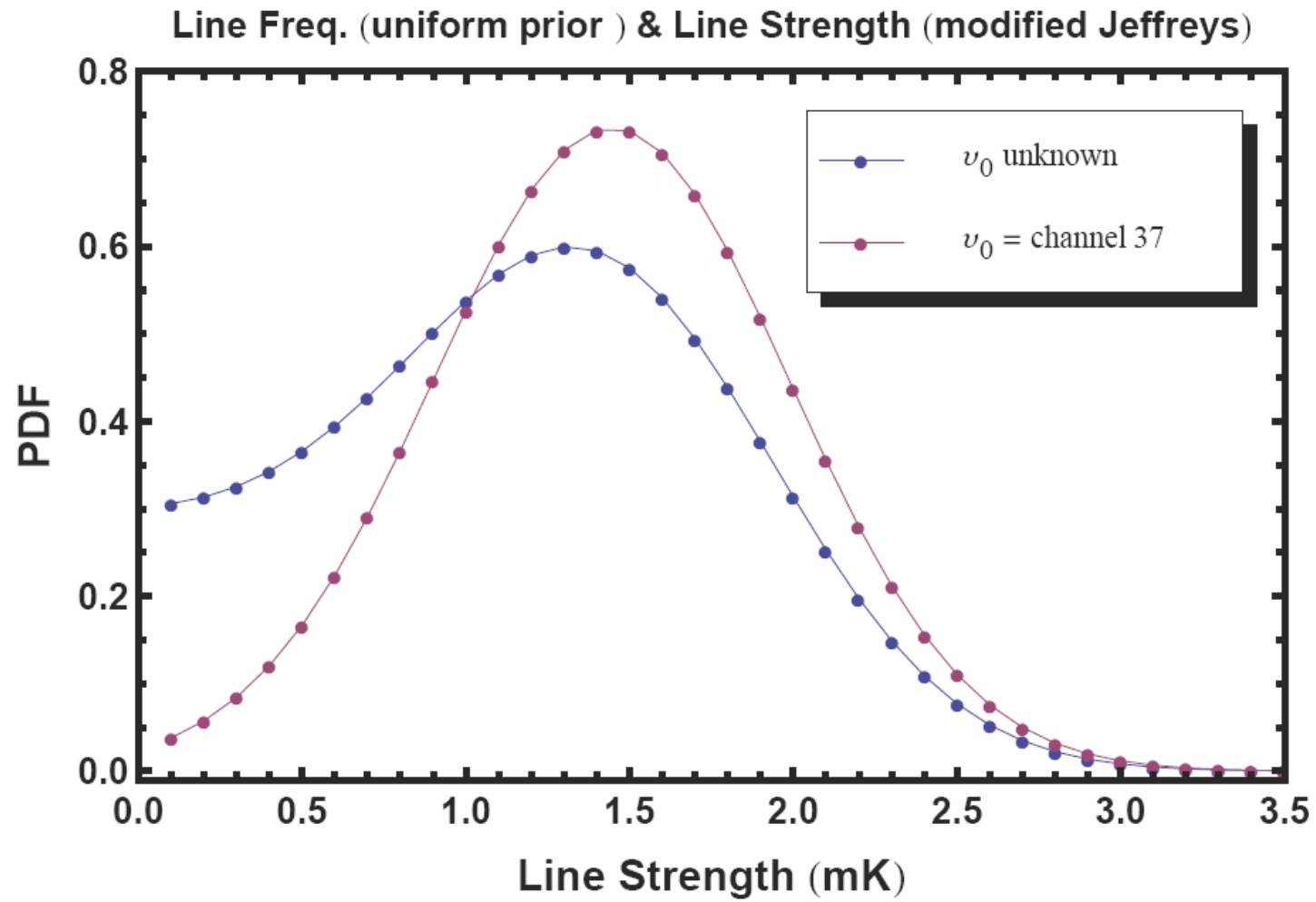
New Bayes factor = 1.0 ,assuming a uniform prior for ν_0 and a Jeffreys prior for T.

Built into any Bayesian model selection calculation is an automatic and quantitative Occam's razor that penalizes more complicated models. One can identify an Occam factor for each parameter that is marginalized. The size of any individual Occam factor depends on our prior ignorance in the particular parameter.

Marginal probability density function for line strength



Marginal probability density function for line strength



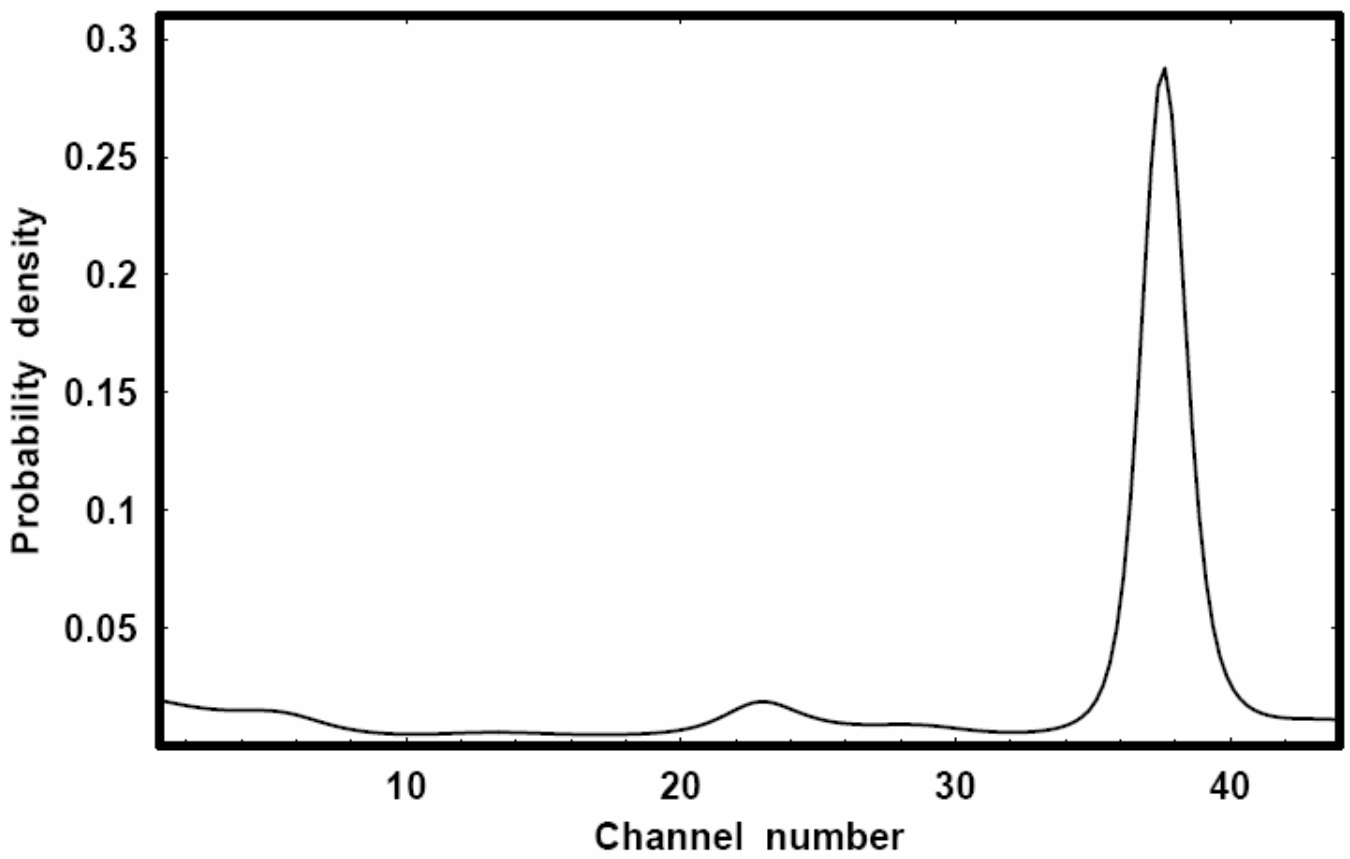
Jeffreys prior

$$p(T \mid M_1, I) = \frac{1}{T} \times \frac{1}{\ln\left[\frac{T_{Max}}{T_{Min}}\right]}$$

Modified Jeffreys prior

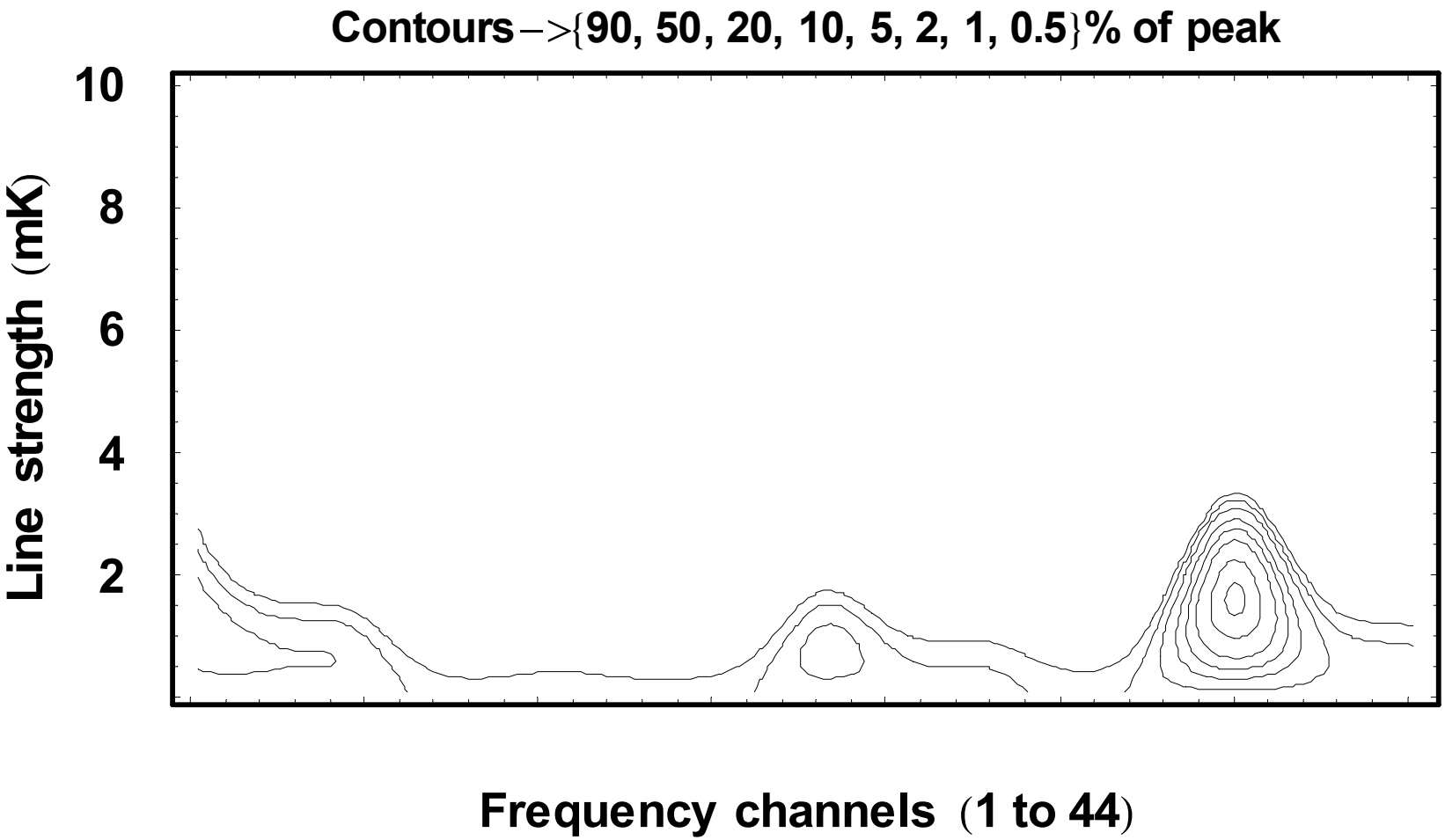
$$p(T \mid M_1, I) = \frac{1}{T + T_0} \times \frac{1}{\ln\left[1 + \frac{T_{Max}}{T_0}\right]}$$

Marginal PDF for line center frequency



Marginal posterior PDF for the line frequency, where the line frequency is expressed as a spectrometer channel number.

Joint probability density function



Generalizations

$$d_i = T \times \exp\left(-\frac{(\nu_i - \nu_0)^2}{2\,\sigma_L^2}\right) + e_i$$

current model

More generally we can write

$$d_i = f_i + e_i$$

where $$f_i = \sum_{\alpha=1}^{m} A_\alpha\, g_\alpha\,(x_i)$$

specifies a linear model with m basis functions $g_\alpha\,(x_i)$

or $$f_i = \sum_{\alpha=1}^{m} A_\alpha\, g_\alpha\,(x_i \mid \theta)$$

specifies a model with m basis functions with an additional set of nonlinear parameters represented by θ.