3

The how-to of Bayesian inference

3.1 Overview

The first part of this chapter is devoted to a brief description of the methods and terminology employed in Bayesian inference and can be read as a stand-alone introduction on how to do Bayesian analysis.[1] Following a review of the basics in Section 3.2, we consider the two main inference problems: parameter estimation and model selection. This includes how to specify credible regions for parameters and how to eliminate nuisance parameters through marginalization. We also learn that Bayesian model comparison has a built-in "Occam's razor," which automatically penalizes complicated models, assigning them large probabilities only if the complexity of the data justifies the additional complication of the model. We also learn how this penalty arises through marginalization and depends both on the number of parameters and the prior ranges of these parameters.

We illustrate these features with a detailed analysis of a toy spectral line problem and in the process introduce the Jeffreys prior and learn how different choices of priors affect our conclusions. We also have a look at a general argument for selecting priors for location and scale parameters in the early phases of an investigation when our state of ignorance is very high. The final section illustrates how Bayesian analysis provides valuable new insights on systematic errors and how to deal with them.

I recommend that Sections 3.2 to 3.5 of this chapter be read twice; once quickly, and again after seeing these ideas applied in the detailed example treated in Sections 3.6 to 3.11.

3.2 Basics

In Bayesian inference, the viability of each member of a set of rival hypotheses, $\{H_i\}$, is assessed in the light of some observed data, D, by calculating the probability of each hypothesis, given the data and any prior information, I, we may have regarding the

[1] The treatment of this topic is a revised version of Section 2 of a paper by Gregory and Loredo (1992), which is reproduced here with the permission of the Astrophysical Journal.

hypotheses and data. Following a notation introduced by Jeffreys (1961), we write such a probability as $p(H_i|D, I)$, explicitly denoting the prior information by the proposition, I, to the right of the bar. At the very least, the prior information must specify the class of alternative hypotheses being considered (hypothesis space of interest), and the relationship between the hypotheses and the data (the statistical model).

The basic rules for manipulating Bayesian probabilities are the sum rule,

$$p(H_i|I) + p(\overline{H}_i|I) = 1, \tag{3.1}$$

and the product rule,

$$\begin{aligned} p(H_i, D|I) &= p(H_i|I)p(D|H_i, I) \\ &= p(D|I)p(H_i|D, I). \end{aligned} \tag{3.2}$$

The various symbols appearing as arguments should be understood as propositions; for example, D might be the proposition, "N photons were counted in a time T." The symbol $\overline{H}_i$ signifies the negation of H_i (a proposition that is true if one of the alternatives to H_i is true), and (H_i, D) signifies the logical conjunction of H_i and D (a proposition that is true only if H_i *and* D are both true). The rules hold for any propositions, not just those indicated above.

Throughout this work, we will be concerned with exclusive hypotheses, so that if one particular hypothesis is true, all others are false. For such hypotheses, we saw in Section 2.5.3 that the sum and product rules imply the generalized sum rule,

$$p(H_i + H_j|I) = p(H_i|I) + p(H_j|I). \tag{3.3}$$

To say that the hypothesis space of interest consists of n mutually exclusive hypotheses means that for the purpose of the present analysis, we are assuming that one of them is true and the objective is to assign a probability to each hypothesis in this space, based on D, I. We will use normalized prior probability distributions, unless otherwise stated, such that

$$\sum_i p(H_i|I) = 1. \tag{3.4}$$

Here a "+" within a probability symbol stands for logical disjunction, so that $H_i + H_j$ is a proposition that is true if either H_i *or* H_j is true.

One of the most important calculating rules in Bayesian inference is Bayes' theorem, found by equating the two right hand sides of Equation (3.2) and solving for $p(H_i|D, I)$:

$$p(H_i|D, I) = \frac{p(H_i|I)p(D|H_i, I)}{p(D|I)}. \tag{3.5}$$

Bayes' theorem describes a type of learning: how the probability for each member of a class of hypotheses should be modified on obtaining new information, D. The probabilities for the hypotheses in the absence of D are called their *prior probabilities*, $p(H_i|I)$, and those including the information D are called their *posterior probabilities*, $p(H_i|D,I)$. The quantity $p(D|H_i,I)$ is called the *sampling probability* for D, or the *likelihood* of H_i, and the quantity $p(D|I)$ is called the *prior predictive probability* for D, or the *global likelihood* for the entire class of hypotheses.

All of the rules we have written down so far show how to manipulate known probabilities to find the values of other probabilities. But to be useful in applications, we additionally need rules that assign numerical values or functions to the initial *direct probabilities* that will be manipulated. For example, to use Bayes' theorem, we need to know the values of the three probabilities on the right side of Equation (3.5). These three probabilities are not independent. The quantity $p(D|I)$ must satisfy the requirement that the sum of the posterior probabilities over the hypothesis space of interest is equal to 1.

$$\sum_i p(H_i|D,I) = \frac{\sum_i p(H_i|I)p(D|H_i,I)}{p(D|I)} = 1. \tag{3.6}$$

Therefore,

$$p(D|I) = \sum_i p(H_i|I)p(D|H_i,I). \tag{3.7}$$

That is, the denominator of Bayes' theorem, which does not depend on H_i, must be equal to the sum of the numerator over H_i. It thus plays the role of a normalization constant.

3.3 Parameter estimation

We frequently deal with problems in which a particular model is assumed to be true and the hypothesis space of interest concerns the values of the model parameters. For example, in a straight line model, the two parameters are the intercept and slope. We can look at this problem as a hypothesis space that is labeled, not by discrete numbers, but by the possible values of two continuous parameters. In such cases, the quantity of interest (see also Section 1.3.2) is a *probability density function* or PDF. More generally, 'PDF' is an abbreviation for a probability distribution function which can apply to both discrete and continuous parameters. For example, given some prior information, M, specifying a parameterized model with one parameter, θ, $p(\theta|M)$ is the prior density for θ, which means that $p(\theta|M)d\theta$ is the prior probability that the true value of the parameter is in the interval $[\theta, \theta + d\theta]$. We use the same symbol, $p(\ldots)$, for probabilities and PDFs; the nature of the argument will identify which use is intended.

Bayes' theorem, and all the other rules just discussed, hold for PDFs, with all sums replaced by integrals. For example, the global likelihood for model M can be calculated with the continuous counterpart of Equation (3.7),

$$p(D|M) = \int d\theta \; p(\theta|M)p(D|\theta, M) = \mathcal{L}(M). \tag{3.8}$$

In words, the global likelihood of a model is equal to the weighted average likelihood for its parameters. We will utilize the global likelihood of a model in Section 3.5 where we deal with model comparison and Occam's razor.

If there is more than one parameter, multiple integrals are used. If the prior density and the likelihood are assigned directly, the global likelihood is an uninteresting normalization constant. The posterior PDF for the parameters is simply proportional to the product of the prior and the likelihood.

The use of Bayes' theorem to determine what one can learn about the values of parameters from data is called *parameter estimation*, though strictly speaking, Bayesian inference does not provide estimates for parameters. Rather, the Bayesian solution to the parameter estimation problem is the full posterior PDF, $p(\theta|D, M)$, and not just a single point in parameter space. Of course, it is useful to summarize this distribution for textual, graphical, or tabular display in terms of a "best-fit" value and "error bars." Possible summaries of the best-fit values are the *posterior mode* (most probable value of θ) or the *posterior mean*,

$$\langle\theta\rangle = \int d\theta \; \theta \; p(\theta|D, M). \tag{3.9}$$

If the mode and mean are very different, the posterior PDF is too asymmetric to be adequately summarized by a single estimate. An allowed range for a parameter with probability content C (e.g., $C = 0.95$ or 95%) is provided by a *credible region*, or highest posterior density region, R, defined by

$$\int_R d\theta \; p(\theta|D, M) = C, \tag{3.10}$$

with the posterior density inside R everywhere greater than that outside it. We sometimes speak picturesquely of the region of parameter space that is assigned a large density as the "posterior bubble." In practice, the probability (density function) $p(\theta|D, M)$ is represented by a finite list of values, p_i, representing the probability in discrete intervals of θ.

A simple way to compute the credible region is to sort these probability values in descending order. Then starting with the largest value, add successively smaller p_i values until adding the next value would exceed the desired value of C. At each step keep track of the corresponding θ_i value. The credible region is the range of θ that just

includes all the θ_i values corresponding to the p_i values added. The boundaries of the credible region are obtained by sorting these θ_i values and taking the smallest and largest values.

3.4 Nuisance parameters

Frequently, a parameterized model will have more than one parameter, but we will want to focus attention on a subset of the parameters. For example, we may want to focus on the implications of the data for the frequency of a periodic signal, independent of the signal's amplitude, shape, or phase. Or we may want to focus on the implications of spectral data for the parameters of some line feature, independent of the shape of the background continuum. In such problems, the uninteresting parameters are known as *nuisance parameters*. As always, the full Bayesian inference is the full joint posterior PDF for all of the parameters; but its implications for the parameters of interest can be simply summarized by integrating out the nuisance parameters. Explicitly, if model M has two parameters, θ and ϕ, and we are interested only in θ, then it is a simple consequence of the sum and product rules (see Section 1.5) that,

$$p(\theta|D,M) = \int d\phi \; p(\theta,\phi|D,M). \tag{3.11}$$

For historical reasons, the procedure of integrating out nuisance parameters is called *marginalization*, and $p(\theta|D,M)$ is called the *marginal posterior PDF* for θ. Equation (3.8) for the global likelihood is a special case of marginalization in which *all* of the model parameters are marginalized out of the joint prior distribution, $p(D,\theta|M)$.

The use of marginalization to eliminate nuisance parameters is one of the most important technical advantages of Bayesian inference over standard frequentist statistics. Indeed, the name "nuisance parameters" originated in frequentist statistics because there is no general frequentist method for dealing with such parameters; they are indeed a "nuisance" in frequentist statistics. Marginalization plays a very important role in this work. We will see a detailed example of marginalization in action in Section 3.6.

3.5 Model comparison and Occam's razor

Often, more than one parameterized model will be available to explain a phenomenon, and we will wish to compare them. The models may differ in form or in number of parameters. Use of Bayes' theorem to compare competing models by calculating the probability of each model as a whole is called *model comparison*. Bayesian model comparison has a built-in "Occam's razor:" Bayes' theorem automatically penalizes complicated models, assigning them large probabilities only if the complexity of the

data justifies the additional complication of the model. See Jeffreys and Berger (1992) for a historical account of the connection between Occam's (Ockham's) razor and Bayesian analysis.

Model comparison calculations require the explicit specification of two or more specific alternative models, M_i. We take as our prior information the proposition that one of the models under consideration is true. Symbolically, we might write this as $I = M_1 + M_2 + \cdots + M_N$, where the "+" symbol here stands for disjunction ("or"). Given this information, we can calculate the probability for each model with Bayes' theorem:

$$p(M_i|D, I) = \frac{p(M_i|I)p(D|M_i, I)}{p(D|I)}. \tag{3.12}$$

We recognize $p(D|M_i, I)$ as the global likelihood for model M_i, which we can calculate according to Equation (3.8). The term in the denominator is again a normalization constant, obtained by summing the products of the priors and the global likelihoods of all models being considered. Model comparison is thus completely analogous to parameter estimation: just as the posterior PDF for a parameter is proportional to its prior times its likelihood, so the posterior probability for a model as a whole is proportional to its prior probability times its global likelihood.

It is often useful to consider the ratios of the probabilities of two models, rather than the probabilities directly. The ratio,

$$O_{ij} = p(M_i|D, I)/p(M_j|D, I), \tag{3.13}$$

is called the *odds ratio* in favor of model M_i over model M_j. From Equation (3.12),

$$\begin{aligned} O_{ij} &= \frac{p(M_i|I)}{p(M_j|I)} \frac{p(D|M_i, I)}{p(D|M_j, I)} \\ &\equiv \frac{p(M_i|I)}{p(M_j|I)} B_{ij}, \end{aligned} \tag{3.14}$$

where the first factor is the prior odds ratio, and the second factor is called the *Bayes factor*. Note: the normalization constant in Equation (3.12) drops out of the odds ratio; this can make the odds ratio somewhat easier to work with. The odds ratio is also conceptually useful when one particular model is of special interest. For example, suppose we want to compare a constant rate model with a class of periodic alternatives, and will thus calculate the odds in favor of each alternative over the constant model.

If we have calculated the odds ratios, O_{i1}, in favor of each model over model M_1, we can find the probabilities for each model in terms of these odds ratios as follows:

$$\sum_{i=1}^{N_{\text{mod}}} p(M_i|D, I) = 1, \tag{3.15}$$

where $N_{\rm mod}$ is the total number of models considered. Dividing through by $p(M_1|D,I)$, we have

$$\frac{1}{p(M_1|D,I)} = \sum_{i=1}^{N_{\rm mod}} O_{i1}. \tag{3.16}$$

Comparing Equation (3.16) to the expression for O_{i1}, given by

$$O_{i1} = p(M_i|D,I)/p(M_1|D,I), \tag{3.17}$$

we have the result that

$$p(M_i|D,I) = \frac{O_{i1}}{\sum_{i=1}^{N_{\rm mod}} O_{i1}}, \tag{3.18}$$

where of course $O_{11} = 1$. If there are only two models, the probability of M_2 is given by

$$p(M_2|D,I) = \frac{O_{21}}{1+O_{21}} = \frac{1}{1+\frac{1}{O_{21}}}. \tag{3.19}$$

In this work, we will assume that we have no information leading to a prior preference for one model over another, so the prior odds ratio will be unity, and the odds ratio will equal the Bayes factor, the ratio of global likelihoods. A crucial consequence of the marginalization procedure used to calculate global likelihoods is that the Bayes factor automatically favors simpler models unless the data justify the complexity of more complicated alternatives. This is illustrated by the following simple example.

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 parameters). To calculate the Bayes factor B_{10} in favor of model M_1, we will need to perform the integral in Equation (3.8) to compute $p(D|M_1,I)$, the global likelihood of M_1. To develop our intuition about the Occam penalty, we will carry out a back-of-the-envelope calculation for the Bayes factor. Often the data provide us with more information about parameters than we had without the data, so that the likelihood function, $\mathcal{L}(\theta) = p(D|\theta,M_1,I)$, will be much more "peaked" than the prior, $p(\theta|M_1,I)$. In Figure 3.1 we show a Gaussian-looking likelihood centered at $\hat{\theta}$, the *maximum likelihood* value of θ, together with a flat prior for θ. Let $\Delta\theta$ be the characteristic width of the prior. For a flat prior, we have that

$$\int_{\Delta\theta} d\theta\; p(\theta|M_1,I) = p(\theta|M_1,I)\Delta\theta = 1. \tag{3.20}$$

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

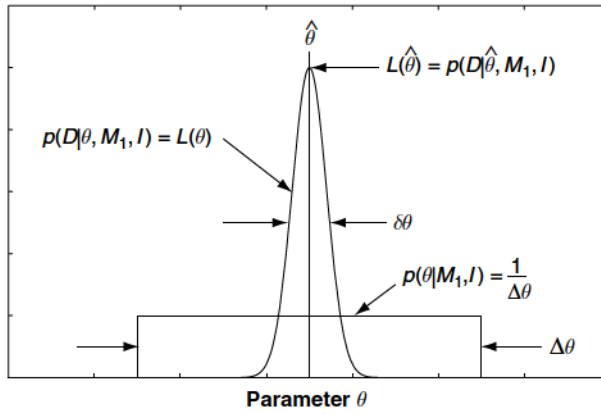

Figure 3.1 The characteristic width $\delta\theta$ of the likelihood peak and $\Delta\theta$ of the prior.

The likelihood has a characteristic width[2] which we represent by $\delta\theta$. The characteristic width is defined by

$$\int_{\Delta\theta} d\theta \; p(D|\theta, M_1, I) = p(D|\hat{\theta}, M_1, I) \times \delta\theta. \tag{3.21}$$

Then we can approximate the global likelihood (Equation (3.8)) for M_1 in the following way:

$$\begin{aligned} p(D|M_1, I) &= \int d\theta \; p(\theta|M_1, I) p(D|\theta, M_1, I) = \mathcal{L}(M_1) \\ &= \frac{1}{\Delta\theta} \int d\theta \; p(D|\theta, M_1, I) \\ &\approx p(D|\hat{\theta}, M_1, I) \frac{\delta\theta}{\Delta\theta} \\ \text{or alternatively,} \quad \mathcal{L}(M_1) &\approx \mathcal{L}(\hat{\theta}) \frac{\delta\theta}{\Delta\theta}. \end{aligned} \tag{3.22}$$

Since model M_0 has no free parameters, no integral need be calculated to find its global likelihood, which is simply equal to the likelihood for model M_1 for $\theta = \theta_0$,

$$p(D|M_0, I) = p(D|\theta_0, M_1, I) = \mathcal{L}(\theta_0). \tag{3.23}$$

Thus the Bayes factor in favor of the more complicated model is

$$B_{10} \approx \frac{p(D|\hat{\theta}, M_1, I)}{p(D|\theta_0, M_1, I)} \frac{\delta\theta}{\Delta\theta} = \frac{\mathcal{L}(\hat{\theta})}{\mathcal{L}(\theta_0)} \frac{\delta\theta}{\Delta\theta}. \tag{3.24}$$

[2] If the likelihood function is really a Gaussian and the prior is flat, it is simple to show that $\delta\theta = \sigma_\theta\sqrt{2\pi}$, where σ_θ is the standard deviation of the posterior PDF for θ.

The 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 second 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 penalty.

Equation (3.22) has the form of the best-fit likelihood times the factor that penalizes M_1. In the above illustrative calculation we assumed a simple Gaussian likelihood function for convenience. In general, the actual likelihood function can be very complicated with several peaks. However, one can always write the global likelihood of a model with parameter θ, as the maximum value of its likelihood times some factor, Ω_θ:

$$p(D|M,I) \equiv \mathcal{L}_{\max}\Omega_\theta. \tag{3.25}$$

The second factor, Ω_θ, is called the *Occam factor* associated with the parameters, θ. It is so named because it corrects the likelihood ratio usually considered in statistical tests in a manner that quantifies the qualitative notion behind "Occam's razor:" simpler explanations are to be preferred unless there is sufficient evidence in favor of more complicated explanations. Bayes' theorem both quantifies such evidence and determines how much additional evidence is "sufficient" through the calculation of global likelihoods.

Suppose M_1 has two parameters θ and ϕ, then following Equation (3.22), we can write

$$\begin{aligned} p(D|M_1,I) &= \iint d\theta d\phi\, p(\theta|M_1,I)p(\phi|M_1,I)p(D|\theta,\phi,M_1,I) \\ &\approx p(D|\hat{\theta},\hat{\phi},M_1,I)\frac{\delta\theta}{\Delta\theta}\frac{\delta\phi}{\Delta\phi} = \mathcal{L}_{\max}\Omega_\theta\Omega_\phi. \end{aligned} \tag{3.26}$$

The above equation assumes independent flat priors for the two parameters. It is clear from Equation (3.26) that the total Occam penalty, $\Omega_{\text{total}} = \Omega_\theta\Omega_\phi$, can become very large. For example, if $\delta\theta/\Delta\theta = \delta\phi/\Delta\phi = 0.01$ then $\Omega_{\text{total}} = 10^{-4}$. Thus for the Bayes factor in Equation (3.24) to favor M_1, the ratio of the maximum likelihoods,

$$\frac{p(D|\hat{\theta},\hat{\phi},M_1,I)}{p(D|M_0,I)} = \frac{\mathcal{L}_{\max}(M_1)}{\mathcal{L}_{\max}(M_0)}$$

must be $\geq 10^4$. Unless the data argue very strongly for the greater complexity of M_1 through the likelihood ratio, the Occam factor will ensure we favor the simpler model. We will explore the Occam factor further in a worked example in Section 3.6.

In the above calculations, we have specifically made a point of identifying the Occam factors and how they arise. In many instances we are not interested in the value of the Occam factor, but only in the final posterior probabilities of the competing models.

Because the Occam factor arises automatically in the marginalization process, its effect will be present in any model selection calculation.

3.6 Sample spectral line problem

In this section, we will illustrate many of the above points in a detailed Bayesian analysis of a toy spectral line problem. In a real problem, as opposed to the hypothetical one discussed below, there could be all sorts of complicated prior information. Although Bayesian analysis can readily handle these complexities, our aim here is to bring out the main features of the Bayesian approach as simply as possible. Be warned; even though it is a relatively simple problem, our detailed solution, together with commentary and a summary of the lessons learned, will occupy quite a few pages.

3.6.1 Background information

In this problem, we suppose that two competing grand unification theories have been proposed. Each one is 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 (background) information and some new data. Both theories make definite predictions in energy ranges beyond the reach of the present generation of particle accelerators. In addition, theory 1 uniquely 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.

Theory 1 also predicts the line will have a Gaussian line shape of the form

$$T\exp\left\{\frac{-(\nu_i-\nu_o)^2}{2\sigma_L^2}\right\} \quad \text{(abbreviated by } Tf_i\text{)}, \tag{3.27}$$

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 channel number and $\nu_o = 37$. The width of the line profile is characterized by σ_L, and $\sigma_L = 2$ channel numbers. The predicted line shape is shown in Figure 3.2.

Data:

To test this prediction, a new spectrometer was mounted on the James Clerk Maxwell telescope on Mauna Kea and the spectrum shown in Figure 3.3 was obtained. The spectrometer has 64 frequency channels with neighboring channels separated by

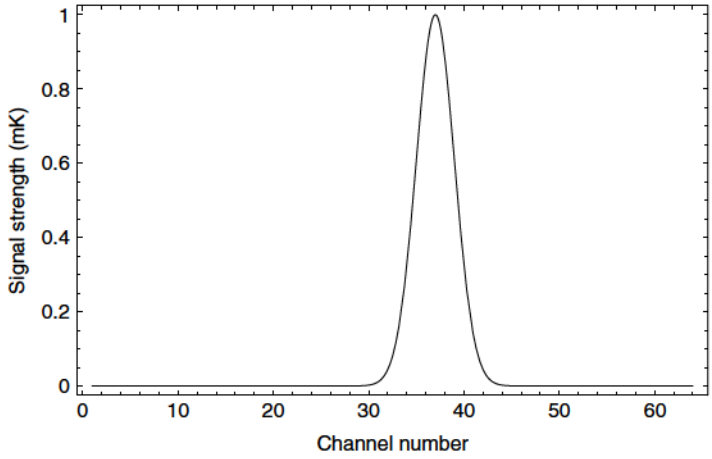

Figure 3.2 Predicted spectral shape according to theory 1.

0.5 σ_L. All channels have Gaussian noise characterized by $\sigma = 1\,\text{mK}$. The noise in separate channels is independent. The data are given in Table 3.1.

Let D be a proposition representing the data from the spectrometer.

$$D \equiv D_1, D_2, \ldots, D_N; \quad N = 64 \tag{3.28}$$

where D_1 is a proposition that asserts that "the data value recorded in the first channel was d_1."

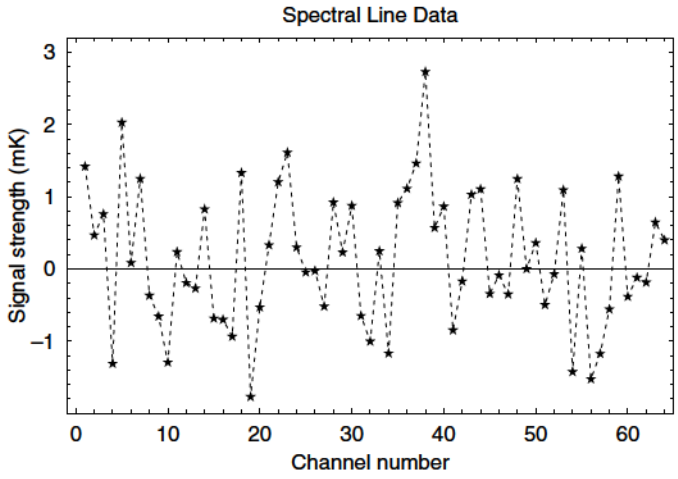

Figure 3.3 Measured spectrum.

Table 3.1 *Spectral line data consisting of 64 frequency channels (#) obtained with a radio astronomy spectrometer. The output voltage from each channel has been calibrated in units of effective black body temperature expressed in mK. The existence of negative values arises from receiver channel noise which gives rise to both positive and negative fluctuations.*

#	mK	#	mK	#	mK	#	mK
1	1.420	17	−0.937	33	0.248	49	0.001
2	0.468	18	1.331	34	−1.169	50	0.360
3	0.762	19	−1.772	35	0.915	51	−0.497
4	−1.312	20	−0.530	36	1.113	52	−0.072
5	2.029	21	0.330	37	1.463	53	1.094
6	0.086	22	1.205	38	2.732	54	−1.425
7	1.249	23	1.613	39	0.571	55	0.283
8	−0.368	24	0.300	40	0.865	56	−1.526
9	−0.657	25	−0.046	41	−0.849	57	−1.174
10	−1.294	26	−0.026	42	−0.171	58	−0.558
11	0.235	27	−0.519	43	1.031	59	1.282
12	−0.192	28	0.924	44	1.105	60	−0.384
13	−0.269	29	0.230	45	−0.344	61	−0.120
14	0.827	30	0.877	46	−0.087	62	−0.187
15	−0.685	31	−0.650	47	−0.351	63	0.646
16	−0.702	32	−1.004	48	1.248	64	0.399

Question: Which theory is more probable?

Based on our current state of information, which includes just the above prior information and the measured spectrum, what do we conclude about the relative probabilities of the two competing theories and what is the posterior PDF for the line strength?

Hypothesis space:

$M_1 \equiv$ "Theory 1 correct, line exists"

$M_2 \equiv$ "Theory 2 correct, no line predicted"

3.7 Odds ratio

To answer the above question, we compute the odds ratio (abbreviated simply by the *odds*) of model M_1 to model M_2.

$$O_{12} = \frac{p(M_1|D,I)}{p(M_2|D,I)}. \tag{3.29}$$

From Equation (3.14) we can write

$$\begin{aligned} O_{12} &= \frac{p(M_1|I)}{p(M_2|I)} \frac{p(D|M_1, I)}{p(D|M_2, I)} \\ &\equiv \frac{p(M_1|I)}{p(M_2|I)} B_{12} \end{aligned} \tag{3.30}$$

where $p(M_1|I)/p(M_2|I)$ is the prior odds, and $p(D|M_1, I)/p(D|M_2, I)$ is the global likelihood ratio, which is also called the Bayes factor.

Based on the prior information given in the statement of the problem, we assign the prior odds $= 1$, so our final odds is given by,

$$O_{12} = \frac{p(D|M_1, I)}{p(D|M_2, I)} \quad \text{(the Bayes factor)}. \tag{3.31}$$

To obtain $p(D|M_1, I)$, the global likelihood of M_1, we need to marginalize over its unknown parameter, T. From Equation (3.8), we can write

$$p(D|M_1, I) = \int dT\, p(T|M_1, I) p(D|M_1, T, I). \tag{3.32}$$

In the following section we will consider what form of prior to use for $p(T|M_1, I)$. In Section 3.7.2 we will show how to evaluate the likelihood, $p(D|M_1, T, I)$.

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

We need to evaluate the global likelihood of model M_1 for use in the Bayes factor. One of the items we need in this calculation is $p(T|M_1, I)$, the prior for T. Choosing a prior is an important part of any Bayesian calculation and we will have a lot to say about this topic in Section 3.10 and other chapters, e.g., Chapter 8, and Sections 9.2.3, 13.3 and 13.4. For this example, we will investigate two common choices: the uniform prior and the Jeffreys prior.[3]

Uniform prior

Suppose we chose a uniform prior for $p(T|M_1, I)$ in the range $T_{\min} \leq T \leq T_{\max}$

$$p(T|M_1, I) = \frac{1}{\Delta T}, \tag{3.33}$$

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 $T_{\max} = 100$ and $T_{\min} = 0.1$. To illustrate the problem, we compare the probability that

[3] If the lower limit on T extended all the way to zero, we would not be able to use a Jeffreys prior because of the infinity at $T = 0$. A modified version of the form, $p(T|M_1, I) = 1/\{(T + a) \ln[(a + T_{\max})/a]\}$, where a is a constant, eliminates this singularity. This *modified Jeffreys* behaves like a uniform prior for $T < a$ and a Jeffreys for $T > a$.

T lies in the upper decade of the prior range (10 to 100 mK) to the lowest decade (0.1 to 1 mK). This is given by

$$\frac{\int_{10}^{100} p(T|M_1, I)dT}{\int_{0.1}^{1} p(T|M_1, I)dT} = 100. \tag{3.34}$$

We see that in this case, a uniform prior implies that the line strength is 100 times more probable to be in the top decade of the predicted range than the bottom, i.e., it is much more probable that T is strong than weak. Usually, expressing great uncertainty in some quantity corresponds more closely to a statement of scale invariance or equal probability per decade. In this situation, we recommend using a Jeffreys prior which is scale invariant.

Jeffreys prior

The form of the prior which represents equal probability per decade (scale invariance) is given by $p(T|M_1, I) = k/T$, where $k =$ constant.

$$\int_{0.1}^{1} p(T|M_1, I)dT = k\int_{0.1}^{1} \frac{dT}{T} = k\ln 10 = \int_{10}^{100} p(T|M_1, I)dT. \tag{3.35}$$

We can evaluate k from the requirement that

$$\int_{T_{\min}}^{T_{\max}} p(T|M_1, I)dT = 1 = k\ln\left(\frac{T_{\max}}{T_{\min}}\right) \tag{3.36}$$

$$\frac{1}{k} = \ln\left(\frac{T_{\max}}{T_{\min}}\right). \tag{3.37}$$

Thus, the form of the Jeffreys prior is given by

$$p(T|M_1, I) = \frac{1}{T\ln(T_{\max}/T_{\min})}. \tag{3.38}$$

A convenient way of summarizing the above comparison between the uniform and Jeffreys prior is to plot the probability of each distribution per logarithmic interval or $p(\ln T|M_1, I)$. This can be obtained from the condition that the probability in the interval T to $T + dT$ must equal the probability in the transformed interval $\ln T$ to $\ln T + d\ln T$.

$$\begin{aligned} p(T|M_1, I)dT &= p(\ln T|M_1, I)d\ln T \\ p(T|M_1, I) &= p(\ln T|M_1, I)\frac{d\ln T}{dT} = \frac{1}{T}p(\ln T|M_1, I) \\ p(\ln T|M_1, I) &= T \times p(T|M_1, I). \end{aligned} \tag{3.39}$$

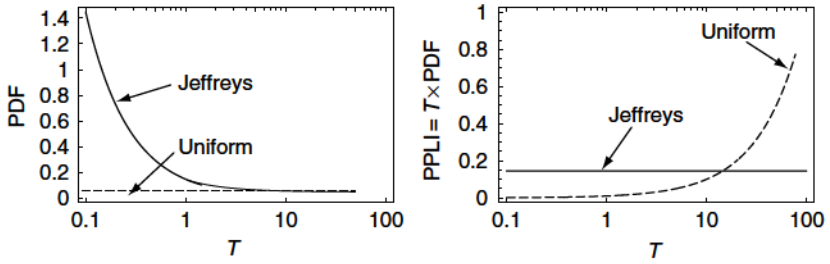

Figure 3.4 The left panel shows the probability density function (PDF), $p(T|M_1, I)$, for the uniform and Jeffreys priors. The right panel shows the probability per logarithmic interval (PPLI), $T \times p(T|M_1, I)$.

Figure 3.4 compares plots of the probability density function (PDF), $p(T|M_1, I)$ (left panel), and the probability per logarithmic interval (PPLI), $T \times p(T|M_1, I)$ (right panel), for the uniform and Jeffreys priors.

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

Let d_i represent the measured data value for the ith channel of the spectrometer. According to model M_1,

$$d_i = Tf_i + e_i, \tag{3.40}$$

where e_i is an error term. Our prior information indicates that this error is caused by receiver noise which has a Gaussian distribution with a standard deviation of σ. Also, from Equation (3.27), we have

$$f_i = \exp\left\{\frac{-(\nu_i - \nu_o)^2}{2\sigma_L^2}\right\}. \tag{3.41}$$

Assuming M_1 is true, then if it were not for the error e_i, d_i would equal Tf_i. Let $E_i \equiv$ "a proposition asserting that the ith error value is in the range e_i to $e_i + de_i$." In this case, we can show (see Section 4.8) that $p(D_i|M_1, T, I) = p(E_i|M_1, T, I)$. If all the E_i are independent[4] then

$$\begin{aligned} p(D|M_1, T, I) &= p(D_1, D_2, \ldots, D_N|M_1, T, I) \\ &= p(E_1, E_2, \ldots, E_N|M_1, T, I) \\ &= p(E_1|M_1, T, I)p(E_2|M_1, T, I)\ldots p(E_N|M_1, T, I) \\ &= \prod_{i=1}^{N} p(E_i|M_1, T, I) \end{aligned} \tag{3.42}$$

[4] We deal with the effect of correlated errors in Section 10.2.2.

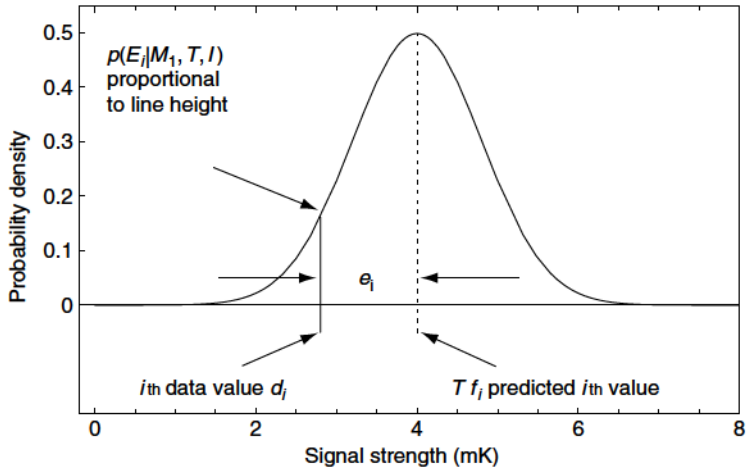

Figure 3.5 Probability of getting a data value d_i a distance e_i away from the predicted value is proportional to the height of the Gaussian error curve at that location.

where $\prod_{i=1}^{N}$ stands for the product of N of these terms. From the prior information, we can write

$$\begin{aligned} p(E_i|M_1, T, I) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{e_i^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(d_i - Tf_i)^2}{2\sigma^2}\right\}. \end{aligned} \tag{3.43}$$

It is apparent that $p(E_i|M_1, T, I)$ is a probability density function since e_i, the value of the error for channel i, is a continuous variable. The factor $(\sigma\sqrt{2\pi})^{-1}$ in the above equation ensures that the integral over e_i from $-\infty$ to $+\infty$ is equal to 1. In Figure 3.5, $p(E_i|M_1, T, I)$ is shown proportional to the height of the Gaussian error curve at the position of the actual data value d_i.

Combining Equations (3.42) and (3.43), we obtain the probability of the entire data set

$$\begin{aligned} p(D|M_1, T, I) &= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(d_i - Tf_i)^2}{2\sigma^2}\right\} \\ &= (2\pi)^{-N/2} \sigma^{-N} \exp\left\{-\frac{\sum_i (d_i - Tf_i)^2}{2\sigma^2}\right\}. \end{aligned} \tag{3.44}$$

In Section 3.7.4, we will need the maximum value of the likelihood given by Equation (3.44). Since we now know all the quantities in Equation (3.44) except T, we can readily compute the likelihood as a function of T in the prior range $0.1 \leq T \leq 100$. The likelihood has a maximum $= 8.520 \times 10^{-37}$ (called the *maximum likelihood*) at $T = 1.561$ mK.

What we want is $p(D|M_1, I)$, the global likelihood of M_1, for use in Equation (3.31). We now evaluate $p(D|M_1, I)$, given by Equation (3.32), for the two different priors discussed in Section 3.7.1, where we argued that the Jeffreys prior matches much more closely the prior information given in this particular problem. Nevertheless, it is interesting to explore what effect the choice of a uniform prior would have on our conclusions. For this reason, we will do the calculations for both priors.

Uniform prior case:

$$\begin{aligned} p(D|M_1, I) &= \frac{(2\pi)^{-N/2}\sigma^{-N}}{\Delta T}\exp\left\{\frac{T\sum d_i f_i}{\sigma^2}\right\}\int_{T_{\min}}^{T_{\max}} dT \exp\left\{\frac{T\sum d_i f_i}{\sigma^2}\right\}\exp\left\{\frac{T\sum d_i f_i}{\sigma^2}\right\} \\ &= 1.131 \times 10^{-38}. \end{aligned} \tag{3.45}$$

According to Equation (3.25), we can always write the global likelihood of a model as the maximum value of its likelihood times an Occam factor, Ω_T, which arises in this case from marginalizing T.

$$\begin{aligned} p(D|M_1, I) &= \mathcal{L}_{\max}(M_1) \times \Omega_T \\ &= \text{maximum value of } [p(D|M_1, T, I)] \times \text{Occam factor} \\ &= 8.520 \times 10^{-37}\,\Omega_T. \end{aligned} \tag{3.46}$$

Comparison of the results of Equations (3.45) and (3.46) leads directly to a value for the Occam factor, associated with our prior uncertainty in the T parameter, of $\Omega_T = 0.0133$.

Jeffreys prior case:

$$\begin{aligned} p(D|M_1, I) &= \frac{(2\pi)^{-N/2}\sigma^{-N}}{\ln(T_{\max}/T_{\min})}\exp\left\{\frac{-\sum d_i^2}{2\sigma^2}\right\} \\ &\quad \times \int_{T_{\min}}^{T_{\max}} dT\, \frac{\exp\left\{\dfrac{T\sum d_i f_i}{\sigma^2}\right\}\exp\left\{-\dfrac{T^2\sum f^{2_i}}{2\sigma^2}\right\}}{T} \\ &= 1.239 \times 10^{-37}. \end{aligned} \tag{3.47}$$

In this case the Occam factor associated with our prior uncertainty in the T parameter, based on a Jeffreys prior, is 0.145. Note: the Occam factor based on the Jeffreys prior is a factor of ≈ 10 less of a penalty than for the uniform prior for the same parameter.

3.7.3 Calculation of $p(D|M_2, I)$

Model M_2 assumes the spectrum is consistent with noise and has no free parameters so in analogy to Equation (3.40), we can write

$$d_i = 0 + e_i \tag{3.48}$$

where e_i = Gaussian noise with a standard deviation of σ. Assuming M_2 is true, then if it were not for the noise e_i, d_i would equal 0.

$$\begin{aligned} p(D|M_2, I) &= (2\pi)^{-N/2}\ \sigma^{-N}\ \exp\left\{-\frac{\sum d_i^2}{2\sigma^2}\right\} \\ &= 1.133 \times 10^{-38}. \end{aligned} \tag{3.49}$$

Since this model has no free parameters, there is no Occam factor, so the global likelihood is also the maximum likelihood, $\mathcal{L}_{\max}(M_2)$, for M_2.

3.7.4 Odds, uniform prior

Substitution of Equations (3.45) and (3.49) into Equation (3.31) leads to an odds ratio for the uniform prior case given by

$$\text{odds} = \frac{1}{\Delta T}\int_{T_{\min}}^{T_{\max}} dT\ \exp\left\{\frac{T\sum d_i f_i}{\sigma^2}\right\}\exp\left\{-\frac{T^2\sum f_i^2}{2\sigma^2}\right\}. \tag{3.50}$$

For $T_{\min} = 0.1$ mK and $T_{\max} = 100$ mK, the odds $= 0.9986$ and

$$p(M_1|D, I) = \frac{1}{1 + \frac{1}{\text{odds}}} = 0.4996. \tag{3.51}$$

Although the ratio of the maximum likelihoods for the two models favors model M_1, by a factor of $\mathcal{L}_{\max}(M_1)/\mathcal{L}_{\max}(M_2) = 8.520 \times 10^{-37}/1.131 \times 10^{-38} \approx 75$, the ratio of the global likelihoods marginally favors M_2 because of the Occam factor which penalizes M_1 for its extra complexity.

3.7.5 Odds, Jeffreys prior

Substitution of Equations (3.47) and (3.49) into Equation (3.31) leads to an odds ratio for the Jeffreys prior case, given by

$$\text{odds} = \frac{1}{\ln(T_{\max}/T_{\min})}\int_{T_{\min}}^{T_{\max}} dT\ \frac{\exp\left\{\frac{T\sum d_i f_i}{\sigma^2}\right\}\ \exp\left\{-\frac{T^2\sum f_i^2}{2\sigma^2}\right\}}{T}. \tag{3.52}$$

For $T_{\min} = 0.1$ mK and $T_{\max} = 100$ mK, the odds $= 10.94$, and $p(M_1|D, I) = 0.916$.

As noted earlier in this chapter, we consider the Jeffreys prior to be much more consistent with the large uncertainty in signal strength which was part of the prior information of the problem. On this basis, we conclude that for our current state of information, $p(M_1|D,I) = 0.916$ and $p(M_2|D,I) = 0.084$.

3.8 Parameter estimation problem

Now that we have solved the model selection problem leading to a significant preference for M_1, which argues for the existence of the short-lived baryon, we would like to compute $p(T|D,M_1,I)$, the posterior PDF for the signal strength. Again we will compute the result for both choices of prior for comparison, but consider the Jeffreys result to be more reasonable for the current problem.

Again, start with Bayes' theorem:

$$\begin{aligned} p(T|D,M_1,I) &= \frac{p(T|M_1,I)p(D|M_1,T,I)}{p(D|M_1,I)} \\ &\propto p(T|M_1,I)p(D|M_1,T,I). \end{aligned} \tag{3.53}$$

We have already evaluated $p(D|M_1,T,I)$ in Equation (3.44). All that remains is to plug in our two different choices for the prior $p(T|M_1,I)$.

Uniform prior case:

$$p(T|D,M_1,I) \propto \exp\left\{\frac{T\sum d_i f_i}{\sigma^2}\right\}\exp\left\{-\frac{T^2\sum f_i^2}{2\sigma^2}\right\} \tag{3.54}$$

Jeffreys prior case:

$$p(T|D,M_1,I) \propto \frac{1}{T}\exp\left\{\frac{T\sum d_i f_i}{\sigma^2}\right\}\exp\left\{-\frac{T^2\sum f_i^2}{2\sigma^2}\right\}. \tag{3.55}$$

Figure 3.6 shows the posterior PDF for the signal strength for both the uniform and Jeffreys priors. As we saw earlier, the uniform prior favors stronger signals.

In our original spectrum, the line strength was comparable to the noise level. How do the results change as we increase the line strength? Figure 3.7 shows a simulated spectrum for a line strength equal to five times the noise σ together with the estimated posterior PDF for the line strength. The increase in line strength has a dramatic effect on the odds which rise to 1.6×10^{12} for the uniform prior and 5.3×10^{12} for the Jeffreys prior.

3.8.1 *Sensitivity of odds to T_{max}*

Figure 3.8 is a plot of the dependence of the odds on the assumed value of $T_{\max}$ for both uniform and Jeffreys priors. We see that under the uniform prior, the odds are

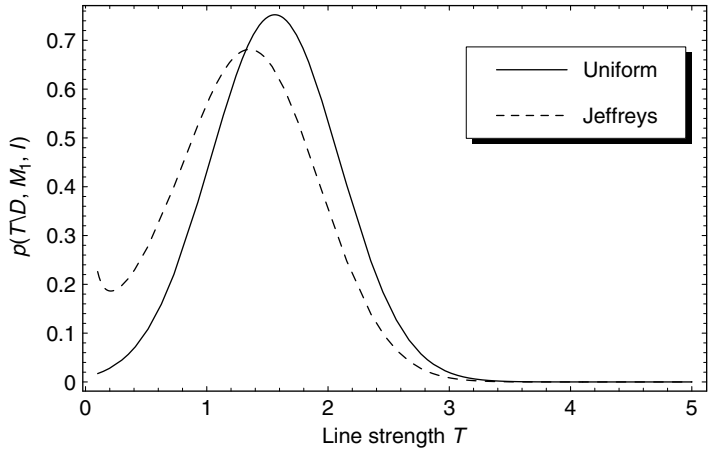

Figure 3.6 Posterior PDF for the line strength, T, for uniform and Jeffreys priors.

much more strongly dependent on the prior range of T than for the Jeffreys case. In both cases, the Occam's razor penalizing M_1 compared to M_2 for its greater complexity increases as the prior range for T increases. Model complexity depends not only on the number of free parameters but also on their prior ranges.

In this problem, we assumed that both the center frequency and line width were accurately predicted by M_1; the only uncertain quantity was the line strength. Suppose the center frequency and/or line width were uncertain as well. In this case, to compute the odds ratio, we would have to marginalize over the prior ranges for these parameters as well, giving rise to additional Occam's factors and a subsequent lowering of the odds. This agrees with our intuition: the more uncertain our prior information about the expected properties of the line, the less significance we attach to any bump in the spectrum.

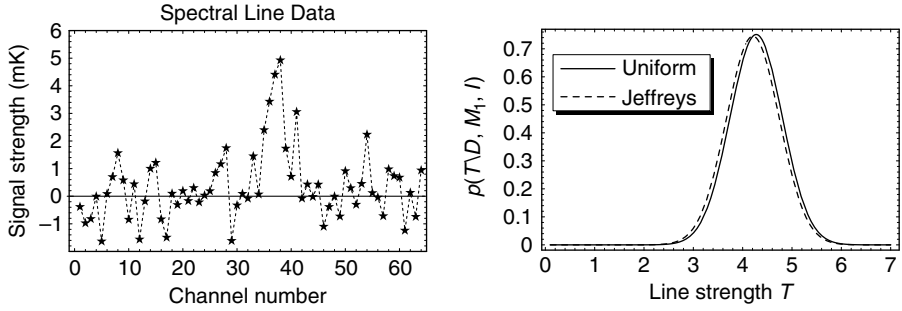

Figure 3.7 The left panel shows a spectrum with a stronger spectral line. The right panel shows the computed posterior PDF for the line strength.

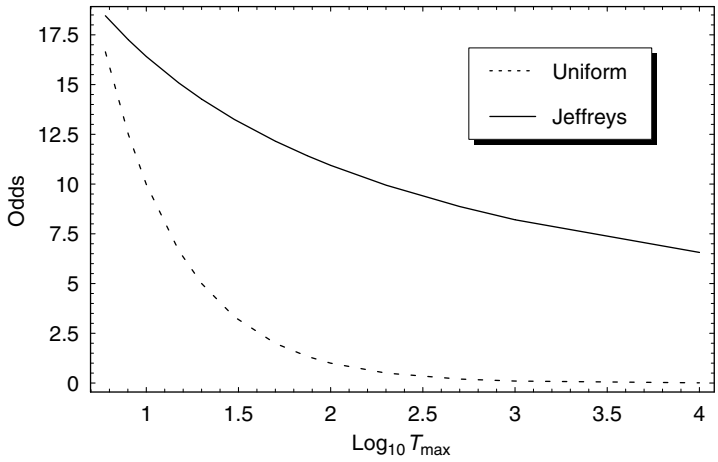

Figure 3.8 The odds ratio versus upper limit on the predicted line strength ($T_{\max}$) for the uniform and Jeffreys priors.

3.9 Lessons

1. In the model selection problem, we are interested in the global probabilities of the two models independent of the most probable model parameters. This was achieved using Bayes' theorem and marginalizing over model M_1's parameter T, the signal strength (model M_2 had no parameters pertaining to the spectral line data). An Occam's razor automatically arises each time a model parameter is marginalized, penalizing the model for prior parameter space that gets ruled out by the data. The larger the prior range that is excluded by the likelihood function, $p(D|M_1, T, I)$, the greater the Occam penalty as can be seen from Figure 3.8. Recall that the global likelihood for a model is the weighted average likelihood for its parameter(s). The weighting function is the prior for the parameter. Thus, the Occam penalty can be very different for two different choices of prior (uniform and Jeffreys). The results are always conditional on the truth of the prior which must be specified in the analysis, and there is a need to seriously examine the consequences of the choice of prior.
2. When the prior range for a parameter spans many orders of magnitude, a uniform prior implies that it is much more probable that the true value of the parameter is in the upper decade. Often, a large prior parameter range can be taken to mean we are ignorant of the scale, i.e., small values of the parameter are equally likely to large values. For these situations, a useful choice is a Jeffreys prior, which corresponds to equal probability per decade (scale invariance). Note: when the range of a prior is a small fraction of the central value, then the conclusion will be the same whether a uniform or Jeffreys prior is used. In the spectrum problem just analyzed, we started out with very crude prior information on the line strength predicted by M_1. Now that we have incorporated the new experimental information D, we have arrived at a posterior probability for the line strength, $p(T|D, M_1, I)$. Were we to obtain more data, D_2, we would set our new prior $p(T|M_1, I_2)$ equal to our current posterior $p(T|D, M_1, I)$, i.e., $I_2 = D, I$. The question of whether to use a Jeffreys or uniform prior would no longer be relevant.

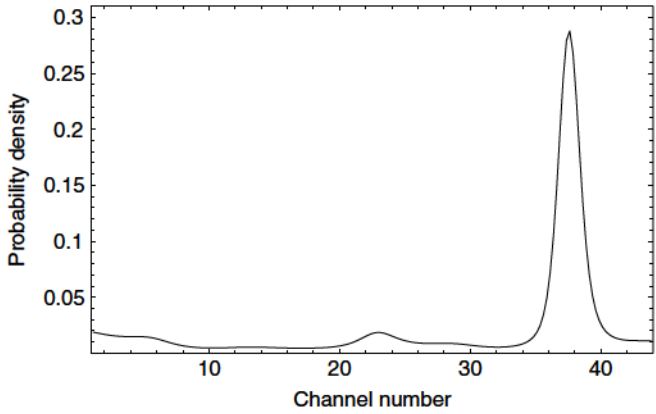

Figure 3.9 Marginal posterior PDF for the line frequency, where the line frequency is expressed as a spectrometer channel number.

3. If the location and line width were also uncertain, we would have to marginalize over these parameters as well, giving rise to other Occam factors which would decrease the odds still further. For example, if the prior range for the expected channel number of the spectral line were increased from less than 1 to 44 channels, the odds would decrease from ≈ 11 to 1, assuming a uniform prior for the line location. We can also compute the marginal posterior PDF for the line frequency for this case which is shown in Figure 3.9. This permits us to update our knowledge of the line frequency given the data and assuming the theory is correct. For further insights on this matter, see the discussion on systematic errors in Section 3.11.
4. Once we established that model M_1 was more probable, we were able to apply Bayes' theorem again, to compute the posterior PDF for the line strength. Note: no Occam factors arise in parameter estimation. Parameter estimation can be viewed as model selection where the competing models all have exactly the same complexity so the Occam penalties are identical and cancel out in the analysis. It can happen that the $p(T|D, M_1, I)$ can be very small for values of T close to zero. One might be tempted to rule out M_2 because it predicts $T = 0$, thus bypassing the model selection problem. This is not wise, however, because the model selection analysis includes Occam factors that could rule out M_1 compared to the simpler M_2. As we noted, these Occam factors do not appear in the parameter estimation analysis.
5. In this toy problem, the spectral line data assume that any background continuum radiation or instrumental DC level has been subtracted off, which can only be done to a certain accuracy. It would be better to parameterize this DC level and marginalize over this parameter so that the effect of our uncertainty in this quantity (see Section 3.11) will be included in our final odds ratio and spectral line parameter estimates. A still more complicated version of this problem is if M_1 simply predicts a certain prior range for the optical depth of the line but

leaves unanswered whether the line will be seen in emission or absorption against the background continuum. In this problem, a Bayesian solution is still possible but will involve a more complicated model of the spectral line data.

3.10 Ignorance priors

In the analysis of the spectral line problem of Section 3.7.1, we considered two different forms of prior (uniform and Jeffreys) for the unknown line temperature parameter. We learned that there was a strong reason for picking the Jeffreys prior in this problem. What motivated a consideration of these particular priors in the first place? In this section we will attempt to answer this question.

As we study any particular phenomenon, our state of knowledge changes. When we are well into the study, our prior for the analysis of new data will be well defined by our previous posterior. But in the earliest phase, our state of "ignorance" will be high. It is therefore useful to have arguments to aid us in selecting an appropriate form of prior to use in such situations. Of course, if we are completely ignorant we cannot even state the problem of interest, and in that case we have no use for a prior. Let us suppose our state of knowledge is sufficient to pose the problem but not much more. For example, we might be interested in the location of the highest point on the equator of Pluto. Are there any general arguments to help us select a suitable prior? In Section 2.6 we saw how to use the *Principle of Indifference* to arrive at a probability distribution for a discrete set of hypotheses.

In the discussion that follows, we will consider a general argument that suggests the form of priors to use for two types of continuous parameters. We will make a distinction between *location parameters*, and *scale parameters*. For example, consider the location of an event in space. To describe this, we must locate the event with respect to some origin and specify the size (scale) of our units of space (e.g., ft, m, light years). The location of an event can be either a positive or negative quantity depending on our choice of origin but the scale (size of our space units) is always a positive quantity. We will first consider a prior for a location parameter.

Suppose we are interested in evaluating $p(X|I)$, where $X \equiv$ "a proposition asserting that the location of the tallest tree along the shore of Lake Superior is between x and $x + dx$." In this statement of the problem, x is measured with respect to a particular survey stake. We will represent the probability density by the function $f(x)$.

What if we consider a different statement of the problem in which the only change is that the origin of our distance measurement has been shifted by an amount c and we are interested in $p(X'|I)$ where $x' = x + c$? If a shift of location (origin) can make the problem appear in any way different, then it must be that we had some kind of prior knowledge about location. In the limit of complete ignorance, the choice of prior would be invariant to a shift in location. Although we are not completely ignorant it still might be useful, in the earliest phase of an investigation, to adopt a prior which is invariant to a shift in location. What form of prior does this imply? If we define our

state of ignorance to mean that the above two statements of the problem are equivalent, then the desideratum of consistency demands that

$$p(X|I)dX = p(X'|I)dX' = p(X'|I)d(X+c) = p(X'|I)dX. \tag{3.56}$$

From this it follows that

$$f(x) = f(x') = f(x+c). \tag{3.57}$$

The solution of this equation is $f(x) = \text{constant}$, so

$$p(X|I) = \text{constant}. \tag{3.58}$$

In the Lake Superior problem, it is apparent that we have knowledge of the upper ($x_{\max}$) and lower ($x_{\min}$) bounds of x, so the constant $= 1/(x_{\max} - x_{\min})$. If we are ignorant of these limits then we refer to $p(X|I)$ as an improper prior, meaning that it is not normalized. An improper prior is useable in parameter estimation problems but is not suitable for model selection problems, because the Occam factors depend on knowing the prior range for each model parameter.

Now consider a problem where we are interested in the mean lifetime of a newly discovered aquatic creature found in the ocean below the ice crust on the moon Europa. We call the lifetime a scale parameter because it can only have positive values, unlike a location parameter which can assume both positive and negative values. Let $\mathcal{T} \equiv$ "the mean lifetime is between τ and $\tau + d\tau$." What form of prior probability density, $p(\mathcal{T}|I)$, should we use in this case? We will represent the probability density by the function $g(\tau)$.

What if we consider a different statement of the problem in which the only change is that the time is measured in units differing by a factor β? Now we are interested in $p(\mathcal{T}'|I)$ where $\tau' = \beta\tau$. If we define our state of ignorance to mean that the two statements of the problems are equivalent, then the desideratum of consistency demands that

$$p(\mathcal{T}|I)d\mathcal{T} = p(\mathcal{T}'|I)d\mathcal{T}' = p(\mathcal{T}'|I)d(\beta\mathcal{T}) = \beta\, p(\mathcal{T}'|I)d\mathcal{T}. \tag{3.59}$$

From this it follows that

$$g(\tau) = \beta g(\tau') = \beta g(\beta\tau). \tag{3.60}$$

The solution of this equation is $g(\tau) = \text{constant}/\tau$, so

$$p(\mathcal{T}|I) = \frac{\text{constant}}{\tau}. \tag{3.61}$$

This form of prior is called the *Jeffreys prior* after Sir Harold Jeffreys who first suggested it. If we have knowledge of the upper ($\tau_{\max}$) and lower ($\tau_{\min}$) bounds of τ then we can evaluate the normalization constant. The result is

$$p(\mathcal{T}|I) = \frac{1}{\tau \ln(\tau_{\max}/\tau_{\min})}. \tag{3.62}$$

Returning to the spectral line problem, we now see another reason for preferring the choice of the Jeffreys prior for the temperature parameter, because it is a scale parameter. In Section 9.2.3, we will discover yet another powerful argument for selecting the Jeffreys prior for a scale parameter.

3.11 Systematic errors

In scientific inference, we encounter at least two general types of uncertainties which are broadly classified as random and systematic. Random uncertainties can be reduced by acquiring and averaging more data. This is the basis behind signal averaging which is discussed in Section 5.11.1. Of course, what appears random for one state of information might later be discovered to have a predictable pattern as our state of information changes.

Some typical examples of systematic errors include errors of calibration of meters and rulers,[5] and stickiness and wear in the moving parts of meters. For example, over time an old wooden meter stick may shrink by as much as a few mm. Some potential systematic errors can be detected by careful analysis of the experiment before performing it and can then be eliminated either by applying suitable corrections or through careful experimental design. The remaining systematic errors can be very subtle, and are detected with certainty only when the same quantity is measured by two or more completely different experimental methods. The systematic errors are then revealed by discrepancies between the measurements made by the different methods.

Bayesian inference provides a powerful way of looking and dealing with some of these subtle systematic errors. We almost always have some prior information about the accuracy of our "ruler." Clearly, if we had no information about its accuracy (in contrast to its repeatability), we would have no logical grounds to use it at all except as a means for ordering events. In this case, we would be expecting no more from our ruler and we would have no concern about a systematic error. What this implies is that we require at least some limited prior information about our ruler's scale to be concerned about a systematic error.

As we have seen, a unique feature of the Bayesian approach is the ability to incorporate prior information and see how it affects our conclusions. In the case of the ruler accuracy, the approach taken is to introduce the scale of the ruler into the calculation as a parameter, i.e., we parameterize the systematic error. We can then treat this as a nuisance parameter and marginalize (integrate over) this parameter to obtain our final inference about the quantity of interest. If the uncertainty in the accuracy of our scale is very large, this will be reflected quantitatively in a larger uncertainty in our final inference.

In a complex measurement, many different types of systematic errors can occur, which in principle, can be parameterized and marginalized. For example, consider the

[5] One important ruler in astronomy is the Hubble relation relating redshift or velocity to distance.

following modification to the spectral line problem of Section 3.6. Even if we know the predicted frequency of the spectral line accurately, the observed frequency depends on the velocity of the source with respect to the observer through the Doppler effect. The observed frequency of the line, f_o, is related to the emitted frequency, f_e by

$$f_o = f_e\left(1 + \frac{v}{c}\right) \quad \text{for } \frac{v}{c} \ll 1, \tag{3.63}$$

where v is the line of sight component of the velocity of the line emitting region and c equals the velocity of light. In our search for a spectral line, we may be examining a small portion of the Orion nebula and only know the distribution of velocities for the integrated emission from the whole nebula, which may be dominated by turbulent and rotational motion of its parts. The unknown factor v introduces a systematic error in our frequency scale. In this case, we might choose to parameterize the systematic error in v by a Gaussian with a mean and σ equal to that of the Orion nebula as a whole.

From the Bayesian viewpoint, we can even consider uncertain scales that arise in a theoretical model as introducing a systematic error on the same footing, for the purposes of inference, as those associated with a measurement. In the above example, we may know the velocity of the source accurately but the theory may be imprecise with regard to its frequency scale.

Of course, the exact form by which we parameterize a systematic error is constrained by our available information, and just as our theories of nature are updated as our state of knowledge changes, so in general will our understanding of these systematic errors.

It is often the case that we can obtain useful information about a systematic error from the interaction between measurements and theory in Bayesian inference. In particular, we can compute the marginal posterior for the parameter characterizing our systematic error as was done in Figure 3.9. This and other points raised in this section are brought out by the problems at the end of this chapter.

The effect of marginalizing over any parameter, whether or not it is associated with a systematic error, is to introduce an Occam factor which penalizes the model for any prior parameter space that gets ruled out by the data through the likelihood function. The larger the prior range that is excluded by the likelihood function, the greater the Occam penalty. It is thus possible to rule out a valid model by employing an artificially large prior for some systematic error or model parameter. Fortunately, Bayesian inference requires one to specify one's choice of prior so its effect on the conclusions can readily be assessed.

3.11.1 Systematic error example

In 1929, Edwin Hubble found a simple linear relationship between the distance of a galaxy, x, and its recessional velocity, v, of the form $v = H_0 x$, where H_0 is known as *Hubble's constant*. Hubble's constant provides the scale of our ruler for astronomical distance determination. An error in H_0 leads to a systematic error in distance

determination. A modern value of $H_0 = 70 \pm 10\,\mathrm{km\ s^{-1}\ Mpc^{-1}}$. Note: astronomical distances are commonly measured in Mpc (a million parsecs). Suppose a particular galaxy has a measured recessional velocity $v_m = (100 \pm 5) \times 10^3\,\mathrm{km\ s^{-1}}$. Determine the posterior PDF for the distance to the galaxy assuming:

1) A fixed value of $H_0 = 70\,\mathrm{km\ s^{-1}\ Mpc^{-1}}$.
2) We allow for uncertainty in the value of Hubble's constant. We assume a Gaussian probability density function for H_0, of the form

$$p(H_0|I) = k \exp\left\{-\frac{(H_0 - 70)^2}{2 \times 10^2}\right\}, \tag{3.64}$$

where k is a normalization constant.
3) We assume a uniform probability density function for H_0, given by

$$p(H_0|I) = \begin{cases} 1/(90 - 50), & \text{for } 50 \leq H_0 \leq 90 \\ 0, & \text{elsewhere.} \end{cases} \tag{3.65}$$

4) We assume a Jeffreys probability density function for H_0, given by

$$p(H_0|I) = \begin{cases} [H_0 \ln(90/50)]^{-1}, & \text{for } 50 \leq H_0 \leq 90 \\ 0, & \text{elsewhere.} \end{cases} \tag{3.66}$$

As usual, we can write

$$v_m = v_{\mathrm{true}} + e \tag{3.67}$$

where v_{true} is the true recessional velocity and e represents the noise component of the measured velocity, v_m. Assume that the probability density function for e can be described by a Gaussian with mean 0 and $\sigma = 5\,\mathrm{km\ s^{-1}}$. To keep the problem simple, we also assume the error in v is uncorrelated with the uncertainty in H_0.

Through the application of Bayes' theorem, as outlined in earlier sections of this chapter, we can readily evaluate the posterior PDF, $p(x|D, I)$, for the distance to the galaxy. The results for the four cases are given below and plotted in Figure 3.10.

Case 1:

$$\begin{aligned} p(x|D, I) \propto p(x|I)\ p(D|x, I) &= p(x|I) \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{e^2}{2\sigma^2}\right\} \\ &= p(x|I) \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(v_m - v_{\mathrm{true}})^2}{2\sigma^2}\right\} \\ &= p(x|I) \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(v_m - H_0 x)^2}{2\sigma^2}\right\}. \end{aligned} \tag{3.68}$$

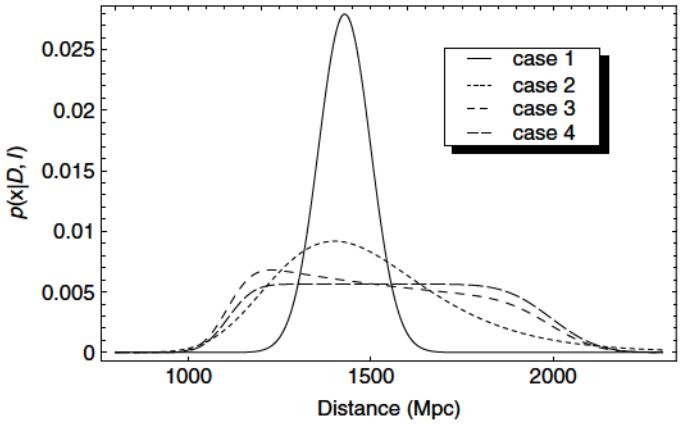

Figure 3.10 Posterior PDF for the galaxy distance, x: 1) assuming a fixed value of Hubble's constant (H_0), 2) incorporating a Gaussian prior uncertainty for H_0, 3) incorporating a uniform prior uncertainty for H_0, and 4) incorporating a Jeffreys prior uncertainty for H_0.

Case 2:

In this case, I incorporates a Gaussian prior uncertainty in the value of H_0.

$$\begin{aligned} p(x|D,I) &= \int_{-\infty}^{\infty} dH_0\ p(x,H_0|D,I) \\ &\propto p(x|I) \int_{-\infty}^{\infty} dH_0\ p(H_0|x,I)\ p(D|x,H_0,I) \\ &= p(x|I) \int_{-\infty}^{\infty} dH_0\ p(H_0|I)\ p(D|x,H_0,I) \\ &= p(x|I) \int_{-\infty}^{\infty} dH_0\ k \exp\left\{-\frac{(H_0-70)^2}{2\times 10^2}\right\} \\ &\quad \times \frac{1}{\sqrt{2\pi}\sigma}\ \exp\left\{-\frac{(v_{\rm m}-H_0 x)^2}{2\sigma^2}\right\}. \end{aligned} \tag{3.69}$$

Case 3:

In this case, I incorporates a uniform prior uncertainty in the value of H_0.

$$\begin{aligned} p(x|D,I) &\propto p(x|I) \int_{50}^{90} dH_0\ p(H_0|I)\ p(D|x,H_0,I) \\ &= p(x|I) \int_{50}^{90} dH_0 \frac{1}{(90-50)} \frac{1}{\sqrt{2\pi}\sigma}\ \exp\left\{-\frac{(v_{\rm m}-H_0 x)^2}{2\sigma^2}\right\}. \end{aligned} \tag{3.70}$$

Case 4:

In this case, I incorporates a Jeffreys prior uncertainty in the value of H_0.

$$\begin{aligned} p(x|D,I) &\propto p(x|I)\int_{50}^{90} dH_0\; p(H_0|I)\; p(D|x,H_0,I) \\ &= p(x|I)\int_{50}^{90} dH_0 \frac{1}{H_0 \ln(90/50)} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(v_m - H_0 x)^2}{2\sigma^2}\right\}. \end{aligned} \tag{3.71}$$

Equations (3.68), (3.69), (3.70), and (3.71) have been evaluated assuming a uniform prior for $p(x|I)$, and are plotted in Figure 3.10. Incorporating the uncertainty in the scale of our astronomical ruler can lead to two effects. Firstly, the posterior PDF for the galaxy distance is broader. Secondly the mean of the PDF is clearly shifted to a larger value. The means of the PDFs for the four cases are 1429, 1486, 1512, and 1556 km s^{-1}, respectively.

It may surprise you that $p(x|D,I)$ becomes asymmetric when we allow for the uncertainty in H_0. One way to appreciate this is to approximate the integral by a weighted summation over a discrete set of choices for H_0. For each choice of H_0, $p(x|D,I)$ is a symmetric Gaussian offset by a distance Δx given by

$$\Delta x = \frac{v_m}{H_0 + \Delta H_0} - \frac{v_m}{H_0} = \left(-\frac{\Delta H_0}{H_0 + \Delta H_0}\right)\frac{v_m}{H_0}. \tag{3.72}$$

For $\Delta H_0 = +20$ km s^{-1} Mpc^{-1}, the bracketed term in Equation (3.72) is equal to -0.22. For $\Delta H_0 = -20$ km s^{-1} Mpc^{-1}, this term is equal to $+0.4$. Thus, the set of discrete Gaussians is more spread out on one side than the other, which accounts for the asymmetry.

3.12 Problems

1. Redo the calculation of the odds for the spectral line problem of Section 3.6 for the case where there is a systematic uncertainty in the line center of ± 5 channels.
2. The prior information is the same as that given for the spectral line problem in Section 3.6 of the text. The measured spectrum is given in Table 3.2. The spectrum consists of 64 frequency channels. Theory predicts the spectral line has a Gaussian shape with a line width $\sigma_L = 2$ frequency channels. The noise in each channel is known to be Gaussian with a $\sigma = 1.0$ mK and the spectrometer output is in units of mK.

   (a) Plot a graph of the raw data.

   (b) Compute the posterior probability of $M_1 \equiv$ "theory 1 is correct, the spectral line exists," for the two cases: (1) Jeffreys prior for the signal strength, and (2) uniform prior. For this part of the problem, assume that the theory predicts that the spectral line is in channel 24. The prior range for the signal strength is 0.1 to 100 mK. In *Mathematica* you can use the command **NIntegrate** to do the numerical integration required in marginalizing over the line strength.

Table 3.2 *Spectral line data consisting of 64 frequency channels obtained with a radio astronomy spectrometer. The output voltage from each channel has been calibrated in units of effective black body temperature expressed in mK. The existence of negative values arises from receiver channel noise which gives rise to both positive and negative fluctuations.*

ch. #	mK	ch. #	mK	ch. #	mK	ch. #	mK
1	0.25	17	−0.42	33	0.44	49	−1.56
2	−0.19	18	1.43	34	0.05	50	−0.64
3	0.25	19	−1.33	35	0.59	51	0.48
4	−0.56	20	0.06	36	0.94	52	1.79
5	−0.41	21	0.82	37	−0.10	53	0.07
6	−0.94	22	0.42	38	0.57	54	1.30
7	0.84	23	3.76	39	0.40	55	0.29
8	−0.30	24	1.10	40	−0.97	56	−0.23
9	−2.06	25	1.31	41	2.20	57	−0.50
10	−1.39	26	1.86	42	0.15	58	0.93
11	0.07	27	0.32	43	−0.37	59	−1.28
12	1.80	28	−1.14	44	−0.67	60	−1.98
13	−1.02	29	1.24	45	−0.05	61	1.85
14	−0.46	30	−0.29	46	−0.20	62	0.89
15	0.29	31	0.02	47	0.65	63	0.65
16	−0.36	32	−1.52	48	−1.24	64	0.28

(c) Explain your reasons for preferring one or the other of the two priors.

(d) On the assumption that the model predicting the spectral line is correct, compute and plot the posterior probability (density function) for the line strength for both priors.

(e) Summarize the posterior probability for the line strength by quoting the most probable value and the (+) and (−) error bars that span the 95% credible region (see the last part of Section 3.3 for a definition of credible region). The credible region can be evaluated by computing the probability for a discrete grid of closely spaced line temperature values. Sort these (probability, temperature) pairs in descending order of probability and then sum the probabilities starting from the highest until they equal 95%. As each term is added, keep track of the upper and lower temperature bounds of the terms included in the sum. *Mathematica* command **Sort[yourdata, OrderedQ[{#2, #1] &];**, will sort the file "yourdata" in descending order according to the first item in each row of the data list.

(f) Repeat the calculations in (b) and (d), only this time, assume that the prior prediction on the location of the spectral line frequency is uncertain; it is predicted to occur somewhere between channels 1 and 50. Assume a uniform

prior for the unknown line center.[6] This will involve computing a two-dimensional likelihood distribution in the variables line frequency and line strength for a discrete set of values of these parameters, and then using a summation operation to approximate integration[7] (you will probably find **NIntegrate** too slow in two dimensions), for marginalizing over both parameters to obtain the global likelihood for computing the odds. For this purpose, you can use a line frequency interval of 1 channel and a signal strength interval of 0.1 mK for 100 intervals. Although this only spans the prior range 0.1 to 10 mK the PDF will be so low beyond 10 mK that it will not contribute significantly to the integral.

(g) Calculate and plot the marginal posterior probabilities for the line frequency.

(h) What additional Occam factor is associated with marginalizing over the prior line frequency range?

3. Plot $p(x|D,I)$ for case 4 (Jeffreys prior) in Section 3.11.1, assuming

$$p(H_0|I) = \begin{cases} \frac{1}{H_0 \ln(80/60)}, & \text{for } 60 \le H_0 \le 80 \\ 0, & \text{elsewhere.} \end{cases} \tag{3.73}$$

Box 3.1

Equation (3.69) can be evaluated using *Mathematica*.
The evaluation will be faster if you compute a **Table** of values for $p(x, H_0|D,I)$ at equally spaced intervals in x, and use **NIntegrate** to integrate over the given range for H_0.

$$p(x|D,I) \propto \textbf{Table}\left[\left\{x, \textbf{NIntegrate}\left[\frac{1}{H_0\sqrt{2\pi}\,\sigma}\ \exp\left(-\frac{(v_m - xH_0)^2}{2\sigma^2}\right), \{H_0, 60, 80\}\right]\right\}, \{x, 800, 2200, 50\}\right]$$

[6] Note: when the frequency range of the prior is a small fraction of center frequency, the conclusion will be the same whether a uniform or Jeffreys prior is assumed for the unknown frequency.

[7] A convenient way to sum elements in a list is to use the *Mathematica* command **Plus@@list**.