Introduction to Bayesian Data Analysis

Phil Gregory

University of British Columbia

March 2010
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)

Introduces statistical inference in the larger context of scientific methods, and includes 55 worked examples and many problem sets.
1. Bayesian primer
2. A simple spectral line problem
   • Background (prior) information
   • Data
   • Hypothesis space of interest
   • Model Selection
   • Choice of prior
   • Likelihood calculation
   • Model selection results
   • Parameter estimation
   • Weak spectral line case
   • Additional parameters
3. Generalizations
4. Occam’s Razor
5. Supplementary: estimating a mean
Bayesian Probability Theory (BPT)  
A primer

BPT = a theory of extended logic

Deductive logic is based on Axiomatic knowledge.

In science we never know any theory of nature is true because our reasoning is based on incomplete information.

Our conclusions are at best probabilities.

Any extension of logic to deal with situations of incomplete information (realm of inductive logic) requires a theory of probability.
A new perception of probability has arisen in recognition that the mathematical rules of probability are not merely rules for manipulating random variables.

They are now recognized as valid principles of logic for conducting inference about any hypothesis of interest.

This view of, ``Probability Theory as Logic'', was championed in the late 20th century by E. T. Jaynes.

“Probability Theory: The Logic of Science”
Cambridge University Press 2003

It is also commonly referred to as Bayesian Probability Theory in recognition of the work of the 18th century English clergyman and Mathematician Thomas Bayes.
Logic is concerned with the truth of propositions. A proposition asserts that something is true.

Examples of propositions:

\[ A \equiv \text{“The newly discovered radio astronomy object is a galaxy.”} \]

\[ B \equiv \text{“The measured redshift of the object is } 0.150 \pm 0.005 \text{.”} \]

\[ A \equiv \text{“Theory } X \text{ is correct.”} \]

\[ \overline{A} \equiv \text{“Theory } X \text{ is not correct.”} \]

\[ A \equiv \text{“The frequency of the signal is between } f \text{ and } f + df \text{.”} \]
We will need to consider compound propositions like $A, B$ which asserts that propositions $A$ and $B$ are true.

$A, B | C$ asserts that propositions $A$ and $B$ are true given that proposition $C$ is true.

**Rules for manipulating probabilities**

**Sum rule:** $p(A | C) + p(A^\complement | C) = 1$

**Product rule:** $p(A, B | C) = p(A | C) p(B | A, C)$

$= p(B | C) p(A | B, C)$

**Bayes theorem:**

$p(A | B, C) = \frac{p(A | C) p(B | A, C)}{p(B | C)}$
How to proceed in a Bayesian analysis?

Write down Bayes’ theorem, identify the terms and solve.

\[
p(H_i \mid D, I) = \frac{p(H_i \mid I) \times p(D \mid H_i, I)}{p(D \mid I)}
\]

Prior probability \quad Likelihood

Posterior probability that \( H_i \) is true, given the new data \( D \) and prior information \( I \)

The likelihood \( p(D \mid H_i, I) \), also written as \( \mathcal{L}(H_i) \), stands for the probability that we would have gotten the data \( D \) that we did, if \( H_i \) is true.
As a theory of extended logic BPT can be used to find **optimal answers** to scientific questions for a given state of knowledge, in contrast to a numerical recipe approach.

**Two basic problems**

1. **Model selection (discrete hypothesis space)**
   “Which one of 2 or more models (hypotheses) is most probable given our current state of knowledge?”
   e.g.
   - Hypothesis $H_0$ asserts that the star has no planets.
   - Hypothesis $H_1$ asserts that the star has 1 planet.
   - Hypothesis $H_i$ asserts that the star has $i$ planets.

2. **Parameter estimation (continuous hypothesis)**
   “Assuming the truth of $H_1$, solve for the probability density distribution for each of the model parameters based on our current state of knowledge.”
   e.g.
   - Hypothesis $H$ asserts that the orbital period is between $P$ and $P+dP$. 
Significance of this development

Probabilities are commonly quantified by a real number between 0 and 1.

The end-points, corresponding to absolutely false and absolutely true, are simply the extreme limits of this infinity of real numbers.

Deductive logic, corresponds to these two extremes of 0 and 1.

Bayesian probability theory spans the whole range.

Deductive logic is just a special case of Bayesian probability theory in the idealized limit of complete information.

End of primer
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

$$T \exp \left\{ -\frac{(\nu_i - \nu_0)^2}{8} \right\} \quad \text{(abbreviated by } T f_i),$$

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

Line profile $f_i$
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:

\[ M_0 \equiv \text{“Model 0, no line exists”} \]
\[ M_1 \equiv \text{“Model 1, line exists”} \]

\( M_1 \) has 1 unknown parameters, the line temperature \( T \).

\( M_0 \) has no unknown parameters.
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 $\sigma = 1 \text{ mK}$. The noise in separate channels is independent. The line center frequency $\nu_0 = 37$. 

*Bayesian Logical Data Analysis for the Physical Sciences* © Cambridge University Press 2005
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

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

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

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

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

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

1. Uniform prior

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

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

There is a problem with this prior if the range of \( T \) is large. In the current example \( T_{\text{min}} = 0.1 \) and \( T_{\text{max}} = 100 \). 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 prior (scale invariant)

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

or equivalently

\[
p(\ln T \mid M_1, I) d\ln T = \frac{d\ln T}{\ln (T_{\text{max}} / T_{\text{min}})}
\]

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

What if the lower bound on \( T \) includes zero? Another alternative is a modified Jeffreys prior of the form.

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

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.
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 $\sigma = 1 \text{ mK}$.

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

\[
p(D|M_1, T, I) = p(D_1, D_2, \ldots, D_N|M_1, T, I) \\
= p(E_1, E_2, \ldots, E_N|M_1, T, I) \\
= p(E_1|M_1, T, I)p(E_2|M_1, T, I)\ldots p(E_N|M_1, T, I) \\
= \prod_{i=1}^{N} p(E_i|M_1, T, I)
\]
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.
From the prior information, we can write

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

\[ = \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{(d_i - Tf_i)^2}{2\sigma^2} \right\} \]

Our final likelihood is given by

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

\[ = (2\pi)^{\frac{-N}{2}} \sigma^{-N} \exp \left\{ -\frac{\sum_i (d_i - Tf_i)^2}{2\sigma^2} \right\} \]

For the given data the maximum value of the likelihood = \(3.80 \times 10^{-37}\)
Final Likelihood $p(D|M_1,T,I)$

$$p(D | M_1, I) = (2\pi)^{-N/2} \sigma^{-N} \text{Exp} \left[-0.5 \sum_{i=1}^{N} \frac{(d_i - T f_i)^2}{\sigma^2} \right]$$

The familiar $\chi^2$ statistic used in least-squares.

Maximizing the likelihood corresponds to minimizing $\chi^2$.

Recall: Bayesian posterior $\propto$ prior $\times$ likelihood.

Thus, only for a uniform prior will a least-squares analysis yield the same solution as the Bayesian posterior.
Maximum and Marginal likelihoods

Our final likelihood is given by

$$p(D \mid M, T, I) = (2\pi)^{-N/2} \sigma^{-N} \exp \left[ -\frac{\sum_i (d_i - Tf_i)^2}{2\sigma^2} \right]$$

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

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

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

Expanded with product rule

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

A uniform prior for $T$ yields $p(D \mid M_1, I) = 5.06 \times 10^{-39}$

A Jeffreys prior for $T$ yields $p(D \mid M_1, I) = 1.74 \times 10^{-38}$
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

$$d_i = 0 + e_i$$

$$p(D|M_0,I) = (2\pi)^{-\frac{N}{2}} \sigma^{-N} \text{Exp} \left[ -\sum_{i=1}^{N} \frac{(d_i - 0)^2}{2 \sigma^2} \right] = 3.26 \times 10^{-51}$$

Model selection results

Bayes factor, uniform prior = $1.6 \times 10^{12}$

Bayes factor, Jeffreys prior = $5.3 \times 10^{12}$

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.
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

$$p(T|D, M_1, I) = \frac{p(T|M_1, I)p(D|M_1, T, I)}{p(D|M_1, I)} \propto p(T|M_1, I)p(D|M_1, T, I)$$
Bayesian Logical Data Analysis for the Physical Sciences © Cambridge University Press 2005
How do our conclusions change when evidence for the line in the data is weaker?

All channels have IID Gaussian noise characterized by $\sigma = 1 \text{ mK}$. The predicted line center frequency $\nu_0 = 37$. 
Weak line model selection results

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

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

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

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 $\nu_0$ 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

New Bayes factor $= 1.0$, assuming a uniform prior for $\nu_0$ 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.
Integration not minimization

In Least-squares analysis we minimize some statistic like $\chi^2$. In a Bayesian analysis we need to integrate.

Parameter estimation: to find the marginal posterior probability density function (PDF) for $T$, we need to integrate the joint posterior over all the other parameters.

$$p(T \mid D, M_1, I) = \int d\nu_0 \ p(T, \nu_0 \mid D, M_1, I)$$

Integartion is more difficult than minimization. However, the Bayesian solution provides the most accurate information about the parameter errors and correlations without the need for any additional calculations, i.e., Monte Carlo simulations.

Next lecture we will discuss an efficient method for Integrating over a large parameter space called Markov chain Monte Carlo (MCMC).
Marginal probability density function for line strength

Line frequency (uniform prior) & line strength (Jeffreys)

\[ n_0 = \text{channel 37} \]

\[ n_0 \text{ unknown} \]
Marginal probability density function for line strength

Line Freq. (uniform prior) & Line Strength (modified Jeffreys)

Jeffreys prior

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

Modified Jeffreys prior

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

Frequency channels (1 to 44)

Contours→(90, 50, 20, 10, 5, 2, 1, 0.5)% of peak
Generalizations

\[
d_i = T \times \exp \left( -\frac{(\nu_i - \nu_0)^2}{2 \sigma^2_L} \right) + e_i
\]

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 \( \theta \).
Generalizations

Examples of linear models

\[ f_i = A_1 + A_2 x + A_3 x^2 + A_3 x^3 + \ldots = \sum_{\alpha=1}^{m} A_{\alpha} g_{\alpha}(x_i) \]

\[ f_i = A_1 \times \exp \left( -\frac{(\nu_i - C_1)^2}{2 \sigma_1^2} \right) \]
where \( C_1 \) and \( \sigma_1 \) are known.

Examples of nonlinear models

\[ f_i = A_1 \cos \omega t + A_2 \sin \omega t \]
where \( A_1, A_2 \) and \( \omega \) are unknowns.

\[ f_i = A_1 \times \exp \left( -\frac{(\nu_i - C_1)^2}{2 \sigma_1^2} \right) + A_2 \times \exp \left( -\frac{(\nu_i - C_2)^2}{2 \sigma_2^2} \right) + \ldots \]
where the \( A' \)s, \( C' \)s and \( \sigma' \)s are unknowns.
Occam’s Razor

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.
Model Comparison and Occam’s Razor

Imagine comparing two models: $M_1$ with a single parameter, $\theta$, and $M_0$ with $\theta$ fixed at some default value $\theta_0$ (so $M_0$ has no free parameter).

We will compare the two models by computing the ratio of their posterior probabilities or Odds.

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

Suppose prior Odds = 1

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

Marginal likelihood for $M_1$

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

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

\[
p(D \mid M_1, I) = \int d\theta \ p(\theta \mid M_1, I) \ p(D \mid \theta, M_1, I)
\]
\[
= \frac{1}{\Delta \theta} \int d\theta \ p(D \mid \theta, M_1, I)
\]

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

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

\[
p(D \mid M_1, I) = p(D \mid \hat{\theta}, M_1, I) \left( \frac{\delta \theta}{\Delta \theta} \right)
\]

Maximum value of the likelihood

Occam factor
Since model $M_0$ has no free parameters the global 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

$$B_{10} = \frac{p(D | M_1, I)}{p(D | M_0, I)} = \frac{p(D | \hat{\theta}, M_1, I)}{p(D | \theta_0 , M_0, I)} \times \frac{\delta \theta}{\Delta \theta}$$
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 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 $\theta$ and $\phi$. Then the Bayes factor would have two Occam factors

$$B_{10} = \frac{p(D | \hat{\theta}, M_1, I)}{p(D | \theta_0, M_0, I)} \frac{\delta \theta}{\Delta \theta} \frac{\delta \phi}{\Delta \phi} = \max \text{ likelihood ratio} \times \Omega_{\theta} \times \Omega_{\phi}$$
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.
Mean: Known noise $\sigma$

The problem is to solve for $p(\mu|D, I)$. The first step is to write down Bayes’ theorem:

$$p(\mu|D, I) = \frac{p(\mu|I) \, p(D|\mu, I)}{p(D|I)}$$

$$p(\mu|I) = K \text{ (constant)} \quad \mu_L \leq \mu \leq \mu_H$$
$$= 0 \quad \text{otherwise.}$$

The likelihood is given by

$$p(D|\mu, I) = \prod_{i=1}^{N} \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{(d_i - \mu)^2}{2\sigma^2} \right\}$$

**Answer**

$$p(\mu|D, I) = \frac{\exp \left\{ -\frac{(\mu - \bar{d})^2}{2\sigma^2/N} \right\}}{\int_{\mu_L}^{\mu_H} \exp \left\{ -\frac{(\mu - \bar{d})^2}{2\sigma^2/N} \right\} d\mu}$$
Mean: Known noise, unequal $\sigma$

\[
p(D|\mu, I) = \prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(d_i-\mu)^2}{2\sigma_i^2}}
\]

\[
= \left[ \prod_{i=1}^{N} \sigma_i^{-1} \right] (2\pi)^{-\frac{N}{2}} \exp \left\{ - \sum_{i=1}^{N} \frac{(d_i - \mu)^2}{2\sigma_i^2} \right\}
\]

\[
= \left[ \prod_{i=1}^{N} \sigma_i^{-1} \right] (2\pi)^{-\frac{N}{2}} \exp \left\{ - \sum_{i=1}^{N} \frac{w_i(d_i - \mu)^2}{2} \right\}
\]

where $w_i = 1/\sigma_i^2$ is called the weight of data value $d_i$.

\[
\text{Answer} \quad p(\mu|D, I) = \frac{\exp \left\{ - \frac{(\mu - \overline{d_w})^2}{2\sigma_w^2} \right\}}{\int_{\mu_L}^{\mu_H} \exp \left\{ - \frac{(\mu - \overline{d_w})^2}{2\sigma_w^2} \right\} d\mu}.
\]

Since the denominator evaluates to a constant, the posterior, within the range $\mu_L$ to $\mu_H$, is simply a Gaussian with variance $\sigma_w^2 = 1/(\sum w_i)$. The most probable value of $\mu$ is the weighted mean $\overline{d_w} = \sum w_i d_i / (\sum w_i)$.

*Bayesian Logical Data Analysis for the Physical Sciences © Cambridge University Press 2005*
It is obvious from the scatter in the measurements compared to the error bars that there is some additional source of uncertainty or the signal strength is variable. For example, additional fluctuations might arise from propagation effects in the interstellar medium between the source and observer.

In the absence of prior information about the distribution of the additional scatter, both the Central Limit Theorem and the Maximum Entropy Principle lead us to adopt a Gaussian distribution because it is the most conservative choice.
Mean: Unknown noise $\sigma$

Now we have two unknowns in our model, $\mu$ and $\sigma$. The joint posterior probability $p(\mu, \sigma|D, I)$ is given by Bayes’ theorem:

$$p(\mu, \sigma|D, I) = \frac{p(\mu, \sigma|I)p(D|\mu, \sigma, I)}{p(D|I)}.$$

We are interested in $p(\mu|D, I)$ regardless of what the true value of $\sigma$ is. In this problem, $\sigma$ is a nuisance parameter so we marginalize over $\sigma$:

$$p(\mu|D, I) = \int p(\mu, \sigma|D, I) d\sigma.$$

From the product rule: $p(\mu, \sigma|I) = p(\mu|I) p(\sigma|\mu, I)$.

Assuming the prior for $\sigma$ is independent of the prior for $\mu$, then

$$p(\mu, \sigma|I) = p(\mu|I) p(\sigma|I).$$
Mean: Unknown noise $\sigma$

We will assume a Jeffreys prior for the scale parameter $\sigma$:

$$
p(\sigma|I) = \begin{cases} 
\frac{K}{\sigma} & \sigma_L \leq \sigma \leq \sigma_H \\
0 & \text{otherwise.}
\end{cases}
$$

The constant $K$ is determined from the condition

$$
\int_{\sigma_L}^{\sigma_H} p(\sigma|I) d\sigma = 1 \quad \Rightarrow \quad K = \frac{1}{\ln(\frac{\sigma_H}{\sigma_L})}
$$

$$
p(\sigma|I) = \frac{1}{\sigma \ln \frac{\sigma_H}{\sigma_L}}.
$$

(9.26)

After some maths and a change of variables we can write

$$
p(\mu|D, I) = \frac{Q^{-\left(\frac{N}{2}\right)} \int_{\tau_L}^{\tau_H} \tau^{\frac{N}{2}-1} e^{-\tau} d\tau}{\int_{\mu_L}^{\mu_H} d\mu \ Q^{-\left(\frac{N}{2}\right)} \int_{\tau_L}^{\tau_H} \tau^{\frac{N}{2}-1} e^{-\tau} d\tau}
$$

$$
\approx \frac{Q^{-\left(\frac{N}{2}\right)}}{\int_{\mu_L}^{\mu_H} d\mu \ Q^{-\left(\frac{N}{2}\right)}},
$$

(9.34)

where

$$
Q = N(\mu - \bar{d})^2 +Nr^2, \quad Nr^2 = \sum(d_i - \bar{d})^2, \quad \text{and} \quad \tau = \frac{Q}{2\sigma^2}
$$

Bayesian Logical Data Analysis for the Physical Sciences © Cambridge University Press 2005
Justification for dropping terms involving \( \int_{\tau_L}^{\tau_H} \tau^{n-1/2} e^{-\tau} d\tau \)

\[
p(\mu|D, I) = \frac{Q^{-\left(\frac{N}{2}\right)}}{\int_{\mu_L}^{\mu_H} d\mu Q^{-\left(\frac{N}{2}\right)} \int_{\tau_L}^{\tau_H} \tau^{\frac{N}{2}-1} e^{-\tau} d\tau} \approx \frac{Q^{-\left(\frac{N}{2}\right)}}{\int_{\mu_L}^{\mu_H} d\mu Q^{-\left(\frac{N}{2}\right)}}
\]

\( N = 10, \; \sigma_H = 5 \; r, \; \sigma_L = 0.5 \; r \)

\[
Q = N (\mu - \bar{d})^2 + N r^2
\]

\[
\tau_H = \frac{Q}{2*\sigma_L^2} \\
\tau_L = \frac{Q}{2*\sigma_H^2}
\]
Mean: Unknown noise $\sigma$

provided $\sigma_L \ll r$ and $\sigma_H \gg r$, where $r$ = the RMS residual of the most probable model fit,

$$p(\mu|D,I) \approx \frac{\left[1 + \frac{(\mu - \bar{d})^2}{r^2}\right]^{-\frac{N}{2}}}{\int_{\mu_L}^{\mu_H} d\mu \left[1 + \frac{(\mu - \bar{d})^2}{r^2}\right]^{-\frac{N}{2}}},$$

where the quantity $Nr^2 = \sum(d_i - \bar{d})^2$
Mean: Unknown noise $\sigma$

Now compare

$$\left[ 1 + \frac{(\mu - \bar{d})^2}{r^2} \right]^{-\frac{N}{2}}$$

with the Student’s $t$ distribution

$$f(t|\nu) = \frac{\Gamma\left[\frac{\nu+1}{2}\right]}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)} \left[ 1 + \frac{t^2}{\nu} \right]^{-\frac{\nu+1}{2}}.$$ (9.36)

If we set

$$\frac{t^2}{\nu} = \frac{(\mu - \bar{d})^2}{r^2},$$

and the number of degrees of freedom $\nu = N - 1$, then equation (9.36) has the same form as the Student’s $t$ distribution.
Fig. Comparison of the computed results for the posterior PDF for the mean radio flux density assuming (a) $\sigma$ known (solid curve), and (b) marginalizing over an unknown $\sigma$ (dashed curve).

Marginalizing over $\sigma$, in effect estimates $\sigma$ from the data. Any variability which is not described by the model is assumed to be noise.

This leads to a broader posterior, $p(\mu | D,I)$, which reflects the larger effective noise.
\[ p(\sigma | D, I) = \frac{p(\sigma | I) \int p(\mu | I) p(D | \mu, \sigma, I) d\mu}{p(D | I)} \]

The distribution is not symmetric like a Gaussian.

\[ (\hat{\sigma}, \langle \sigma \rangle, \sqrt{\langle \sigma^2 \rangle}) \text{ of } p(\sigma | D, I) \]

are all different.

\[ \hat{\sigma} = r \]

\[ r = \text{the RMS deviation from } \bar{d}. \]

and

\[ \langle \sigma^2 \rangle = \frac{1}{N - 1} \sum_{i=1}^{N} (d_i - \bar{d})^2 \]

Compare the latter with the frequentist sample variance,

\[ S^2 = \sum (d_i - \bar{d})^2 / (N - 1) \]
Model selection results assuming noise $\sigma$ unknown

Suppose are prior information indicates that measurement errors are independent and identically distributed (IID) Gaussians with an unknown $\sigma$ in the range 0 to 4. Can we still analyze this case?

Yes, we can treat $\sigma$ as a nuisance parameter with a modified Jeffreys prior extending from $\sigma = 0 \rightarrow 4.0$. In this case the Bayesian analysis automatically inflates $\sigma$ to include anything in the data that cannot be explained by the model.

For the stronger line:
\[
\text{Bayes factor} = 1.15 \times 10^{10} \text{ (Jeffreys prior for } T \text{ and } \sigma )
\]

For the weaker line:
\[
\text{Bayes factor} = 22.6 \text{ (Jeffreys prior for } T \text{ and } \sigma )
\]

Here the Bayes factor has actually increased because the analysis finds the effective noise is $< 1 \text{ mK}$.
Model selection results with allowance for extra noise

What if we allowed for the possibility of an extra Gaussian noise term with unknown sigma = $s$, added in quadrature to the known measurement error.

We treat $s$ as a nuisance parameter with a Jeffreys prior extending from $s=0 \rightarrow 0.5$ times the peak-to-peak spread of the data values. In this case the Bayesian analysis automatically inflates $s$ to include anything in the data that cannot be explained by the model.

For the stronger line:
Bayes factor = $1.24 \times 10^{10}$ (Jeffreys prior for $T$ and $\sigma$)

For the weaker line:
Bayes factor = 9.14 (Jeffreys prior for $T$ and $\sigma$)