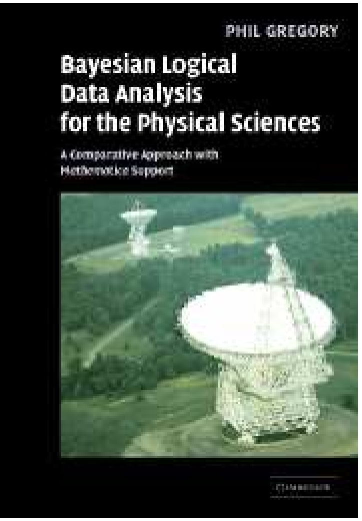
SAMSI Astrostatistics Tutorial

Models with Gaussian Uncertainties (lecture 1)

Phil Gregory

University of British Columbia

2006


PHIL GREGORY
Bayesian Logical Data Analysis for the Physical Sciences
A Comparative Approach with Mathematica Support
CAMBRIDGE

Description
Table of contents
Excerpt
Index
Copyright
Frontmatter

Resources and solutions
This title has free online support material available.
View material...

Details
128 line diagrams 4 half-tones 74 exercises 132 figures 55 worked examples
Weight: 1.146 kg

Bayesian Logical Data Analysis for the Physical Sciences

A Comparative Approach with Mathematica Support

P. C. Gregory (University of British Columbia, Vancouver)

Hardback (ISBN-10: 052184150X | ISBN-13: 9780521841504)
Also available in eBook format

Increasingly, researchers in many branches of science are coming into contact with Bayesian statistics or Bayesian probability theory. By encompassing both inductive and deductive logic, Bayesian analysis can improve model parameter estimates by many orders of magnitude. It provides a simple and unified approach to all data analysis problems, allowing the experimenter to assign probabilities to competing hypotheses of interest, on the basis of the current state of knowledge. This book provides a clear exposition of the underlying concepts with large numbers of worked examples and problem sets. The book also discusses numerical techniques for implementing the Bayesian calculations, including an introduction to Markov Chain Monte-Carlo integration and linear and nonlinear least-squares analysis seen from a Bayesian perspective. In addition, background material is provided in appendices and supporting Mathematica notebooks are available, providing an easy learning route for upper-undergraduates, graduate students, or any serious researcher in physical sciences or engineering.

- **Introduces statistical inference in the larger context of scientific methods, and includes many worked examples and problem sets.**
- **Presents Bayesian theory but also compares and contrasts with other existing ideas.**
- **Mathematica based support software is available.**

outline

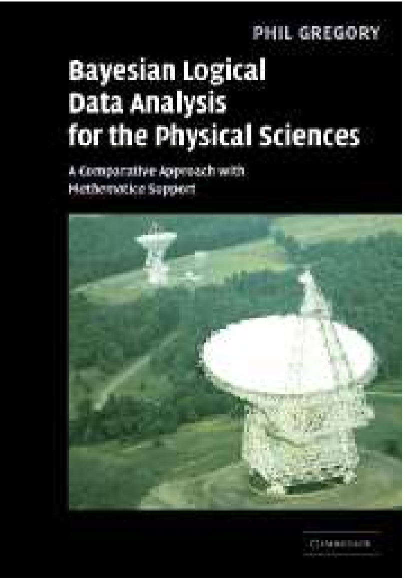
Bayesian Logical Data Analysis for the Physical Sciences

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)

Inference with Gaussian uncertainties

Appendix A. Singular value decomposition
Appendix B. Discrete Fourier Transform
Appendix C. Difference in two samples
Appendix D. Poisson ON/OFF details
Appendix E. Multivariate Gaussian from maximum entropy.

CAMBRIDGE


PHIL GREGORY
Bayesian Logical Data Analysis for the Physical Sciences
A Comparative Approach with Mathematica Support
CAMBRIDGE

Resources and solutions

Book Preface (11 Kb)

Mathematica Tutorial (1415 Kb)

Errata (206 Kb)

Revisions (268 Kb)

Book website: www.cambridge.org/052184150X

Outline

1. Starting point for any Bayesian analysis? 1
2. A simple spectral line problem 2
    - Background (prior) information
    - Data 3
    - Hypothesis space of interest 4
    - Model Selection 5
    - Choice of prior 6
    - Likelihood calculation 7
    - Model selection results 8
    - Parameter estimation 9
    - Weak spectral line case 10
    - Additional parameters 11
3. Generalizations 12
4. Summary 13
5. Supplementary: estimating a mean 14

outline

How to proceed in a Bayesian analysis?

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

Prior probability **Likelihood**

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

Posterior probability that $\mathbf{H_i}$ is true, given the new data D and prior information I

Normalizing constant

Every item to the right of the vertical bar | is assumed to be true

The likelihood p(D| $\mathbf{H_i}$, I), also written as $\mathcal{L}(\mathbf{H_i})$, stands for the probability that we would have gotten the data D that we did, if $\mathbf{H_i}$ is true.

Simple Spectral Line Problem

Background (prior) information:

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

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

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

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

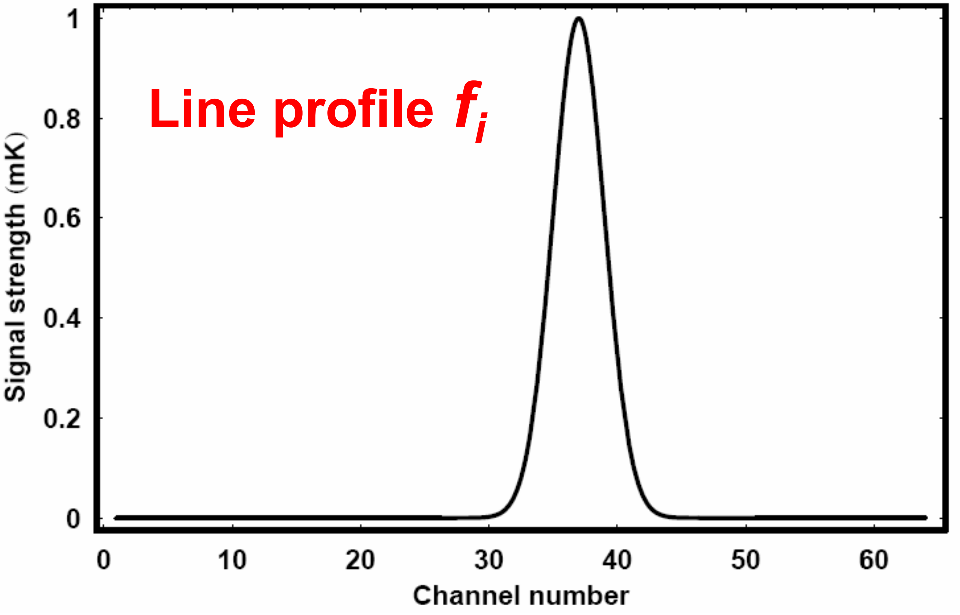
Simple Spectral Line Problem

outline

The predicted line shape has the form

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

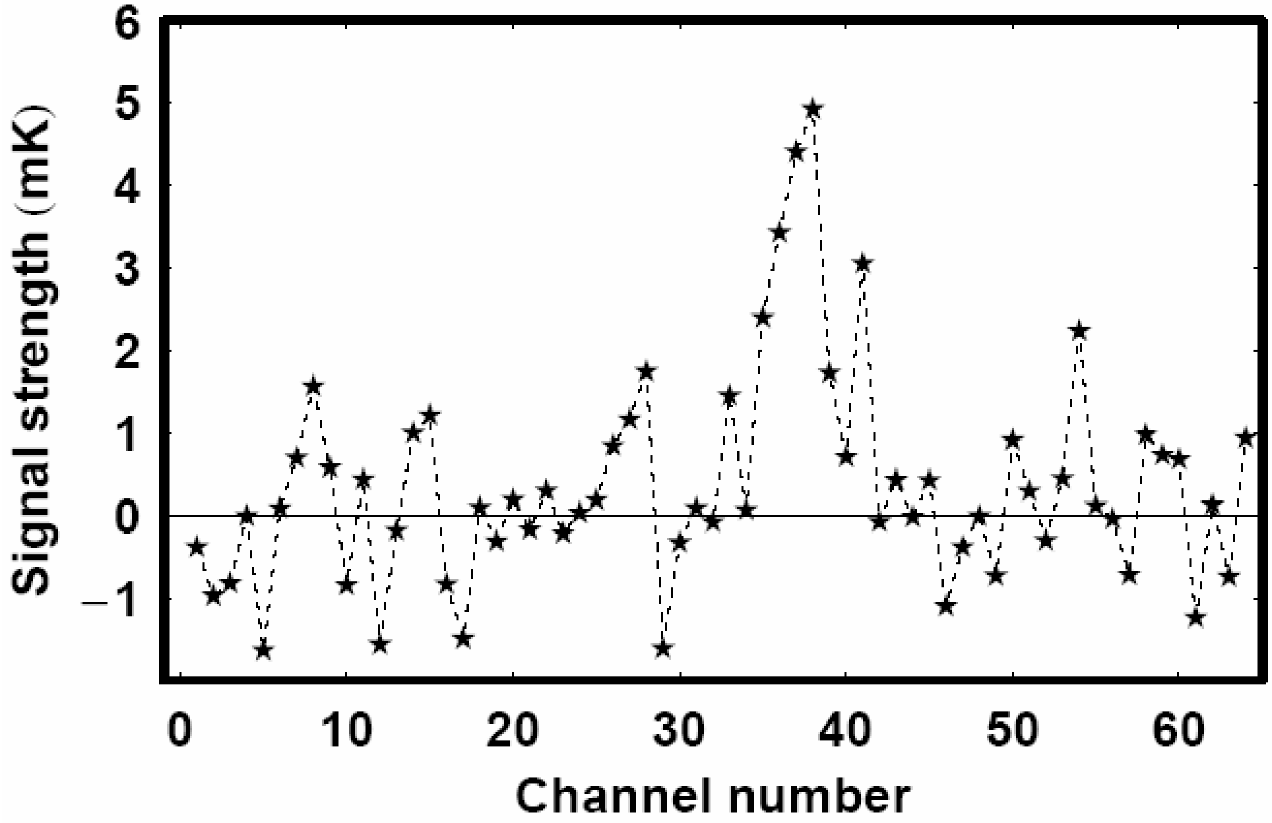
where the signal strength is measured in temperature units of mK and *T* is the amplitude of the line. The frequency, ν_i , is in units of the spectrometer channel number and the line center frequency ν_0 = 37.



Data

outline

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 with neighboring.



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

outline

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 line strength?

Hypothesis space of interest for model selection part:

$$M_1 \equiv \text{“Theory 1 correct, line exists”}$$

$$M_2 \equiv \text{“Theory 2 correct, no line predicted”}$$

outline

Model selection

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

$$O_{12} = \underbrace{\frac{p\,(M_1 \mid D,\ I)}{p\,(M_2 \mid D,\ I)}}_{\text{posterior probability ratio}} = \overbrace{\frac{\dfrac{p\,(M_1|I)\ p\,(D|M_1,\ I)}{p\,(D|I)}}{\dfrac{p\,(M_2|I)\ p\,(D|M_2,\ I)}{p\,(D|I)}}}^{\text{Expand numerator and denominator with Bayes' theorem}} = \underbrace{\frac{p\,(M_1 \mid I)}{p\,(M_2 \mid I)}}_{\text{prior probability ratio}}\ \underbrace{\frac{p\,(D \mid M_1,\ I)}{p\,(D \mid M_2,\ I)}}_{\text{Bayes factor}}$$

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

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

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

Expanded with product rule

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

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

Investigate two common choices

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

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

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

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

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

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

outline

2. Jeffreys prior

(Scale invariant)

$$p(T|M_1, I) = \frac{1}{T \ln\left(\frac{T_{\max}}{T_{\min}}\right)}$$

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

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

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

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

outline

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

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

$$d_i = T\,f_i + e_i \quad \text{and} \quad f_i = \exp\left(\frac{-(\nu_i - \nu_0)^2}{8}\right),$$

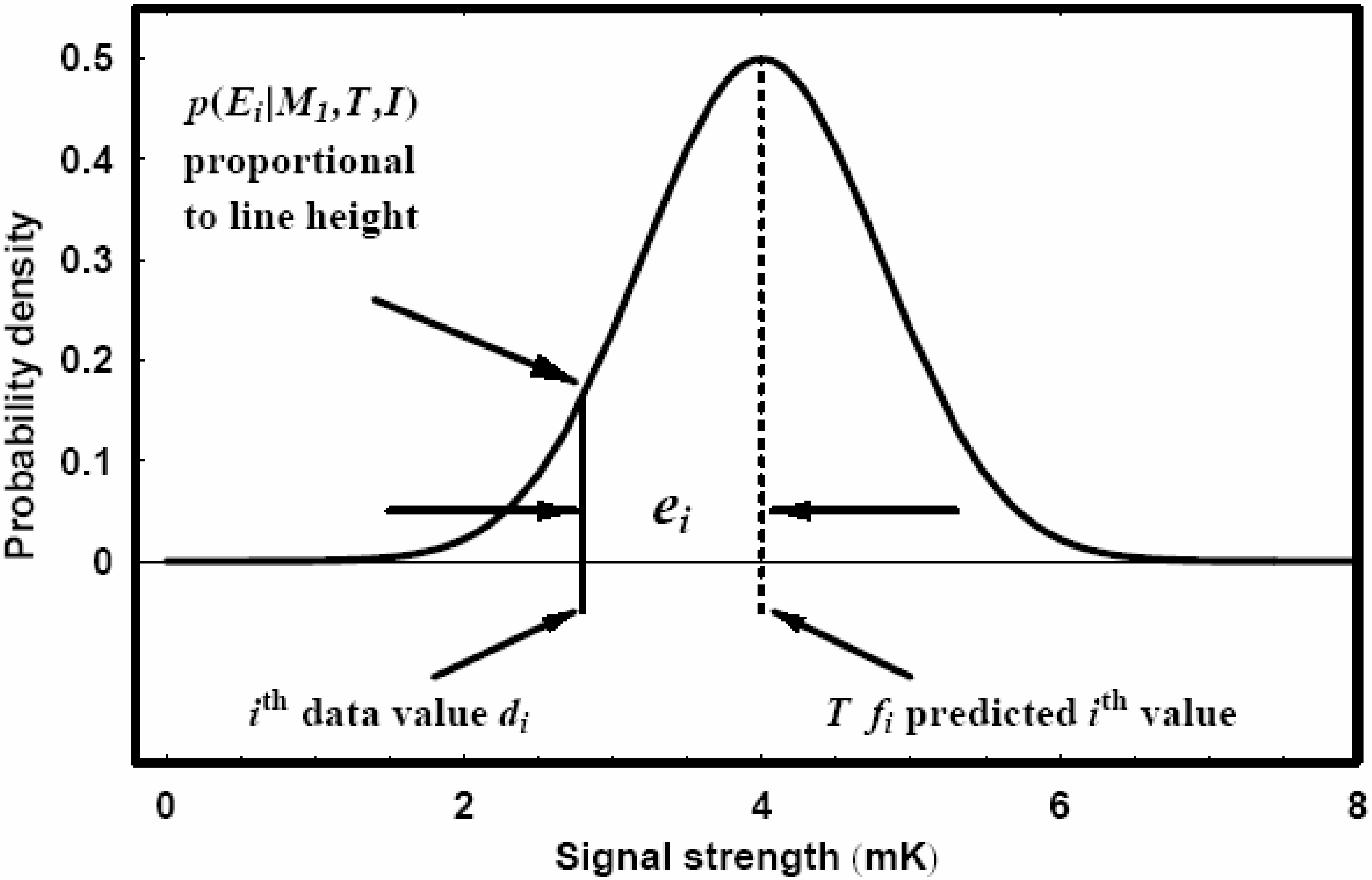
and e_i represents the error component in the measurement. Our prior information indicates that e_i has a Gaussian distribution with a $\sigma = 1\ mK$.

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

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

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

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



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

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

outline

From the prior information, we can write

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

Our final likelihood is given by

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

outline

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

Model M_2 assumes the spectrum is consistent with noise and has no free parameters so we can write

$$d_i = 0 + e_i$$

$$p(D|M_2, I) \quad = \quad (2\pi)^{-\frac{N}{2}} \, \sigma^{-N} \, e^{-\frac{\sum d_i^2}{2\sigma^2}}$$

Model selection results

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

Bayes factor, Jeffreys prior = 5.3 x 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.

outline

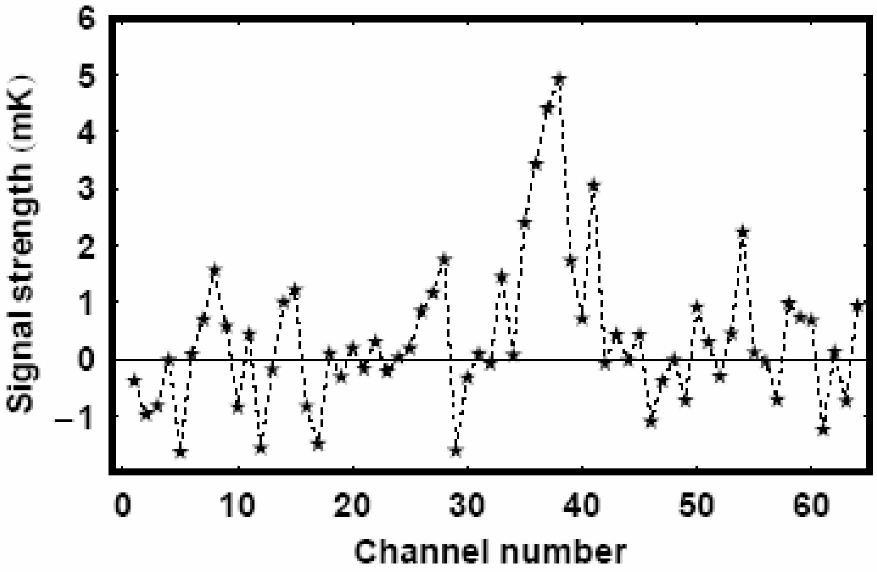
Parameter Estimation Problem

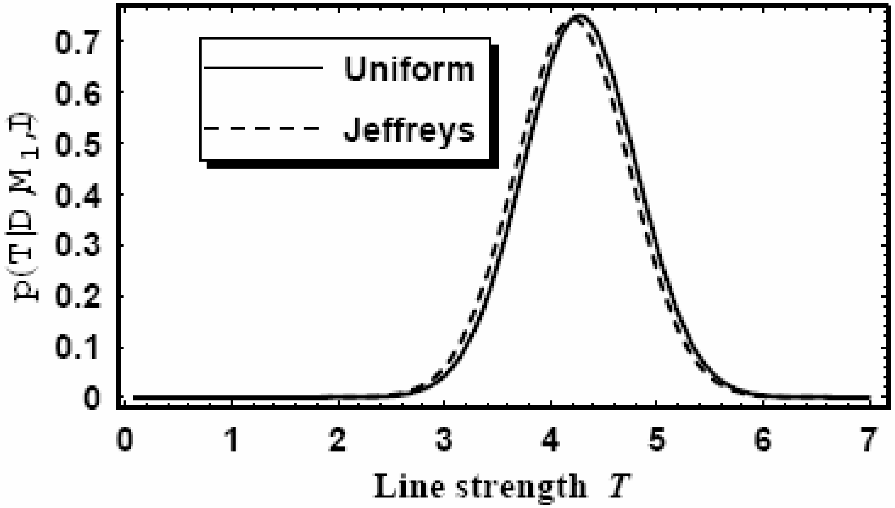
Now that we have solved the model selection problem leading to a significant preference for M_1, we would now like to compute $p(T|D,M_1, I)$, the posterior PDF for the signal strength.

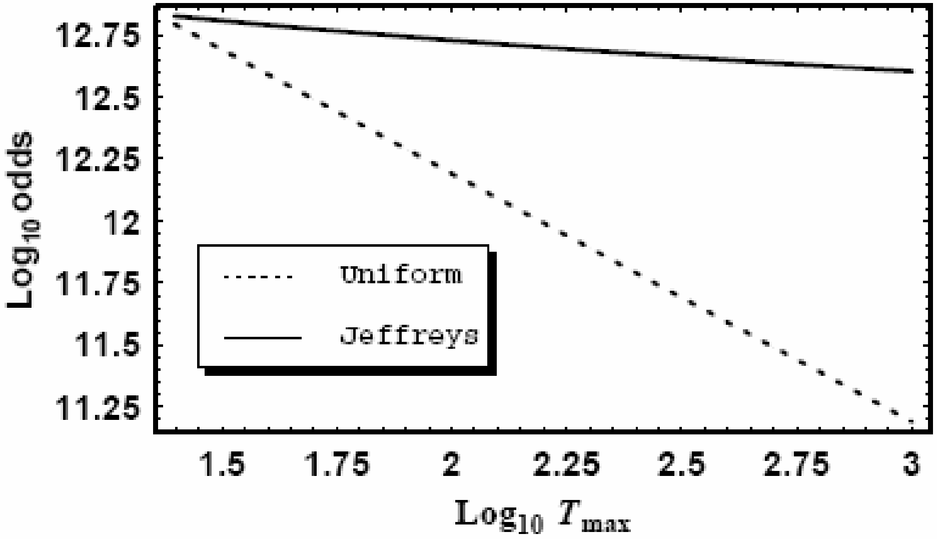
Again, start with Bayes' theorem

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

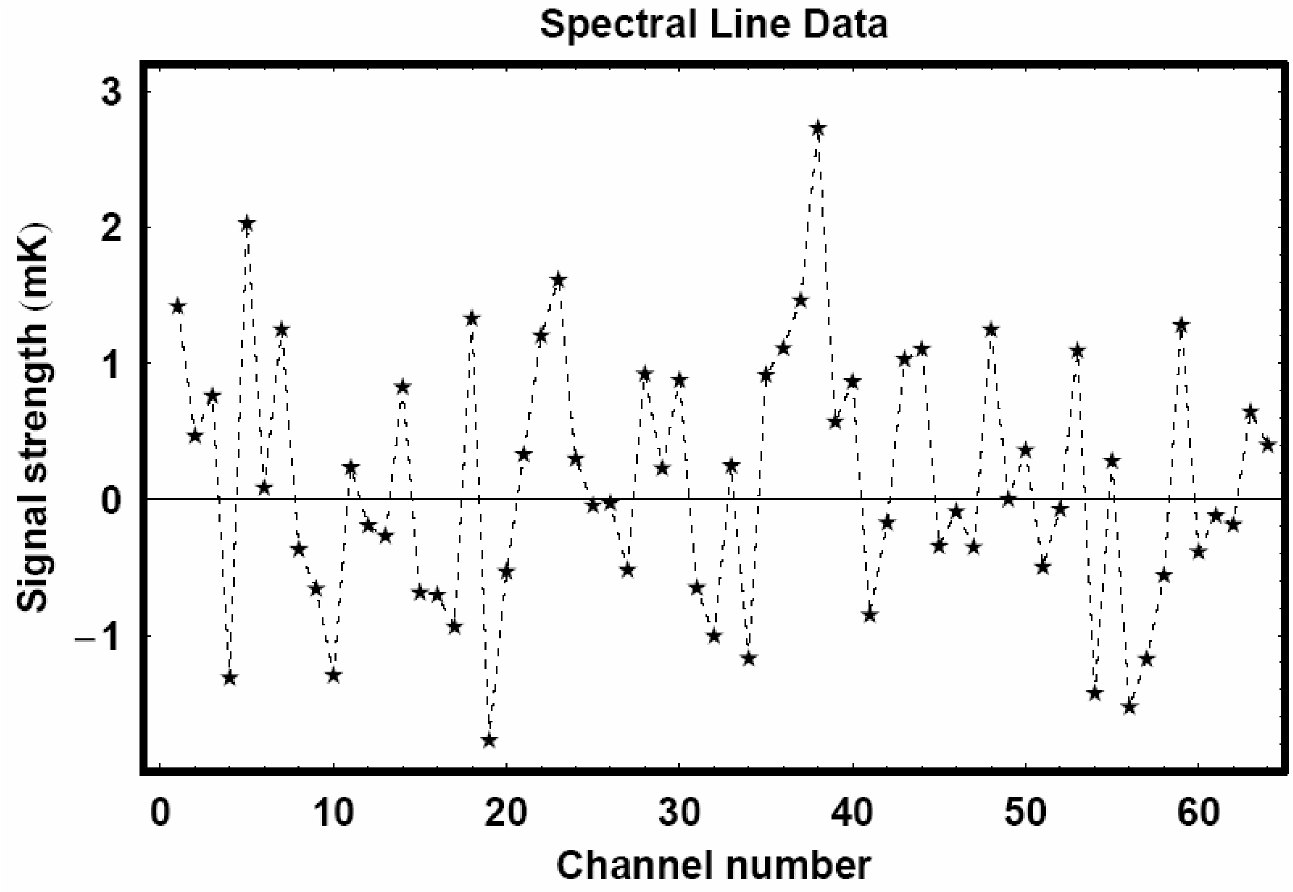
outline


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


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


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

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



All channels have IID Gaussian noise characterized by σ = 1 mK.
The predicted line center frequency ν_0 = 37.

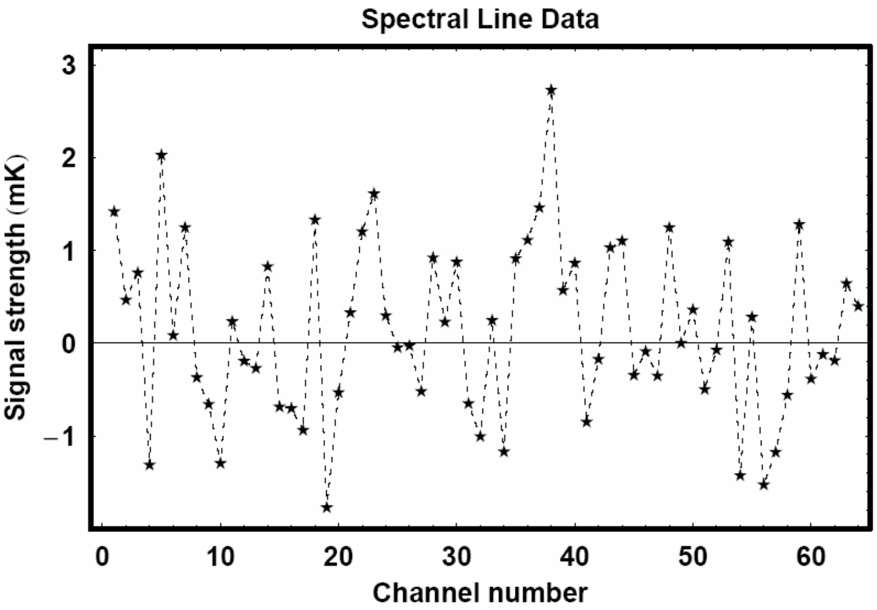
Model selection results

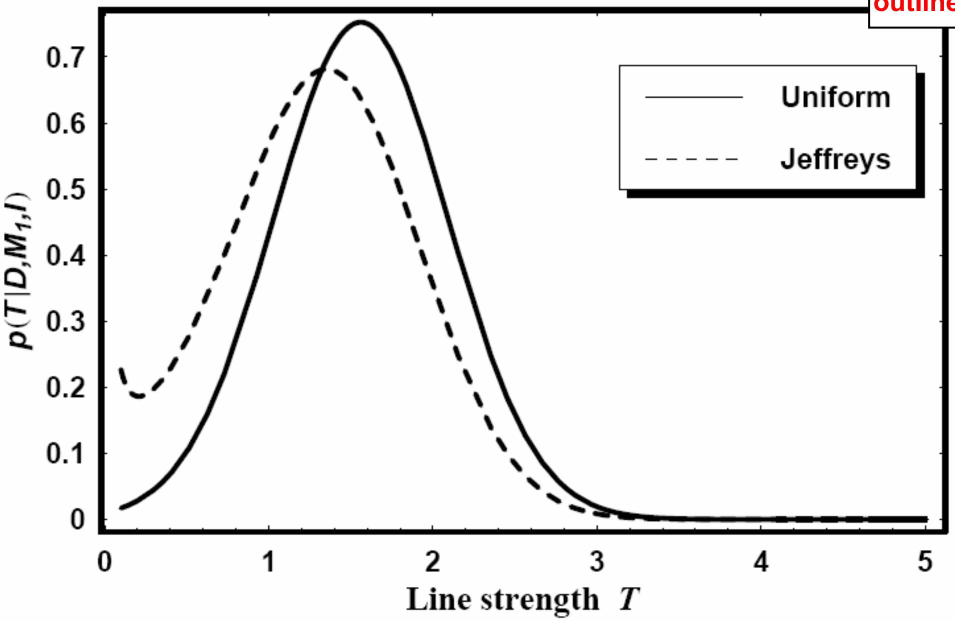
Bayes factor, uniform prior = 1.0

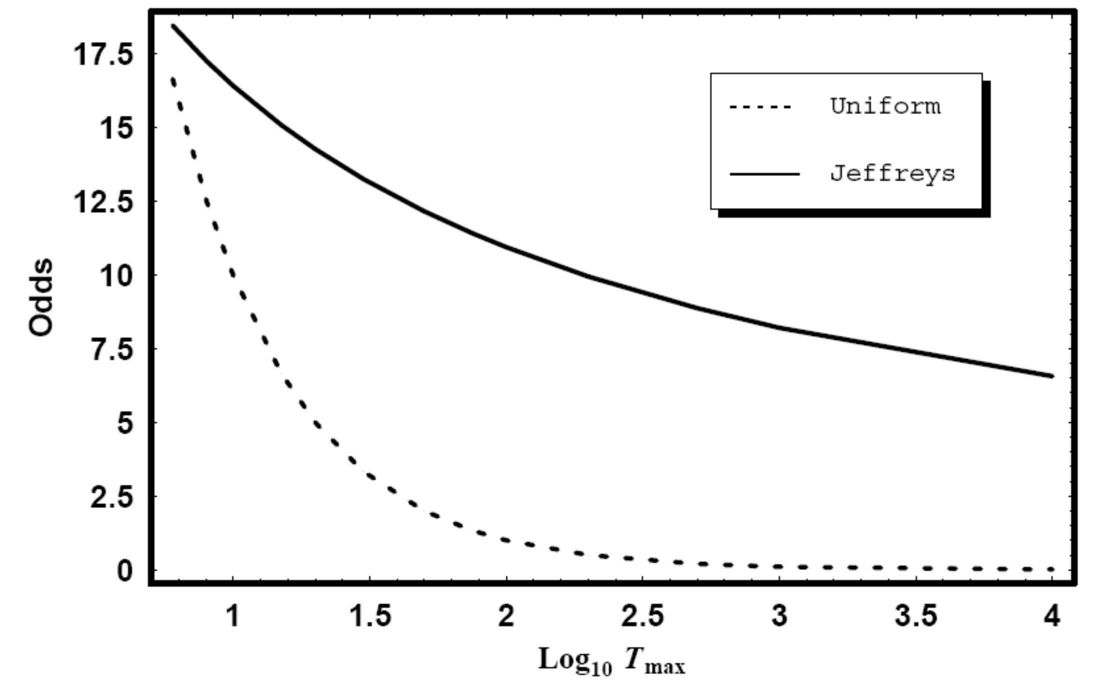
Bayes factor, Jeffreys prior = 11.

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 several different priors.

outline


Spectral Line Data
Signal strength (mK)
3
2
1
0
−1
0
10
20
30
40
50
60
Channel number


$p(T|D,M_1,I)$
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
Line strength T
Uniform
Jeffreys


Odds
17.5
15
12.5
10
7.5
5
2.5
0
1
1.5
2
2.5
3
3.5
4
$\text{Log}_{10}\ T_{\text{max}}$
Uniform
Jeffreys

What if we were uncertain about the line center frequency?

outline

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

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

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

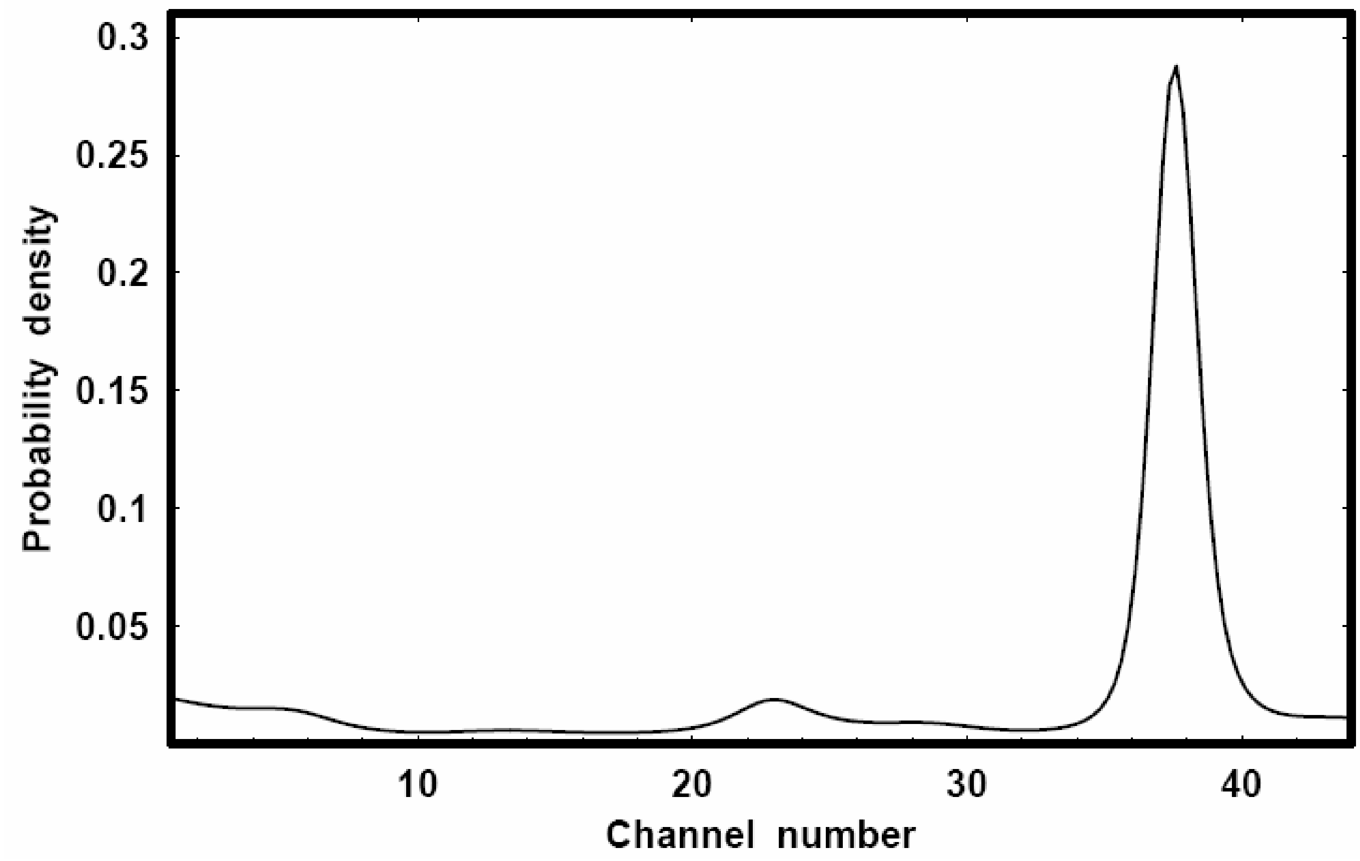
$$= \int_{\nu_0}\int_T p(T \mid M_1, I)\; p(\nu_0 \mid M_1, I)\; p(D \mid M_1, T, \nu_0, I)\, dT\, d\nu_0$$

Assumes independent priors

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

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

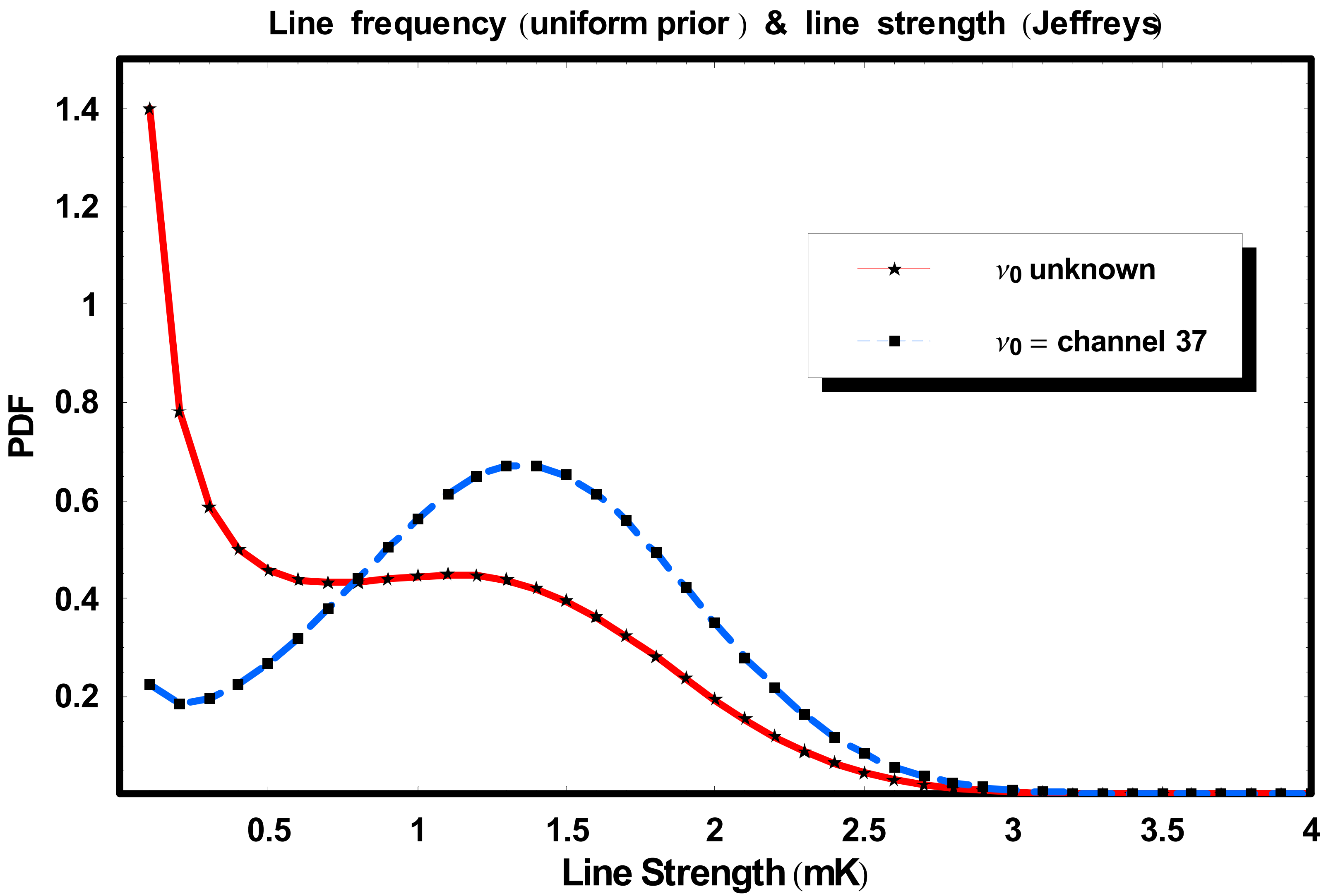
Marginal PDF for line center frequency



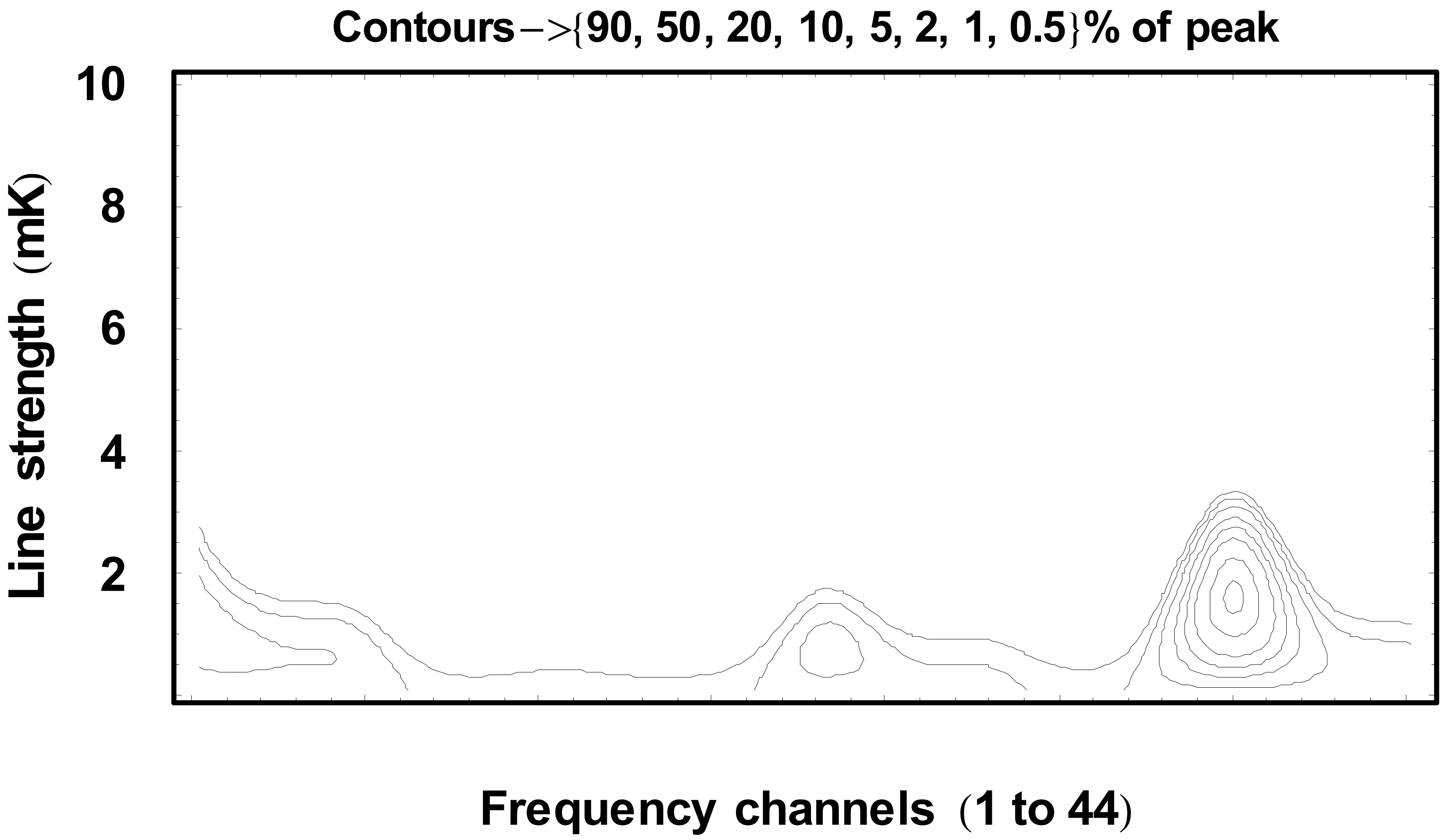
Marginal posterior PDF for the line frequency, where the line frequency is expressed as a spectrometer channel number.

outline

Marginal probability density function for line strength



Joint probability density function



Generalizations

$$\underbrace{d_i = T \times \exp\left(-\frac{(\nu_i - \nu_0)^2}{8}\right)}_{\text{current model}} + e_i$$

More generally we can write

$$d_i = f_i + e_i$$

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

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

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

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

Generalizations

outline

Examples of linear models

$$f_i = A_1 + A_2\, x + A_3\, x^2 + A_3\, x^3 + \;.\,.\,.\,. = \sum_{\alpha=1}^{m} A_\alpha\; g_\alpha\,(x_i)$$

$$f_i = A_1 \times \exp\left(-\frac{(\nu_i - C_1)^2}{2\,\sigma_1^2}\right)$$

where C_1 and σ_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 ω are unknowns.

$$f_i = A_1 \times \exp\left(-\frac{(\nu_i - C_1)^2}{2\,\sigma_1^2}\right) + A_2 \times \exp\left(-\frac{(\nu_i - C_2)^2}{2\,\sigma_2^2}\right) + \,.\,.\,.$$

where the A's, C's and σ's are unknowns.

outline

Data D   **Model M**   **Prior I**

Posterior

Linear models (uniform priors)

Posterior has a single peak (multi-dimensional Gaussian)

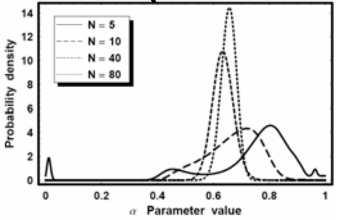
Parameters given by the normal equations of linear least-squares

No integration required very fast using linear algebra

(chapter 10)

Nonlinear models

+ linear models (non-uniform priors)



Posterior may have multiple peaks

Brute force integration

For some nuisance parameters analytic integration sometimes possible

Asymptotic approx.'s

peak finding algorithms (e.g. Levenberg-Marquardt)

Laplace approx.'s

(chapter 11)

Moderate dimensions

quadrature

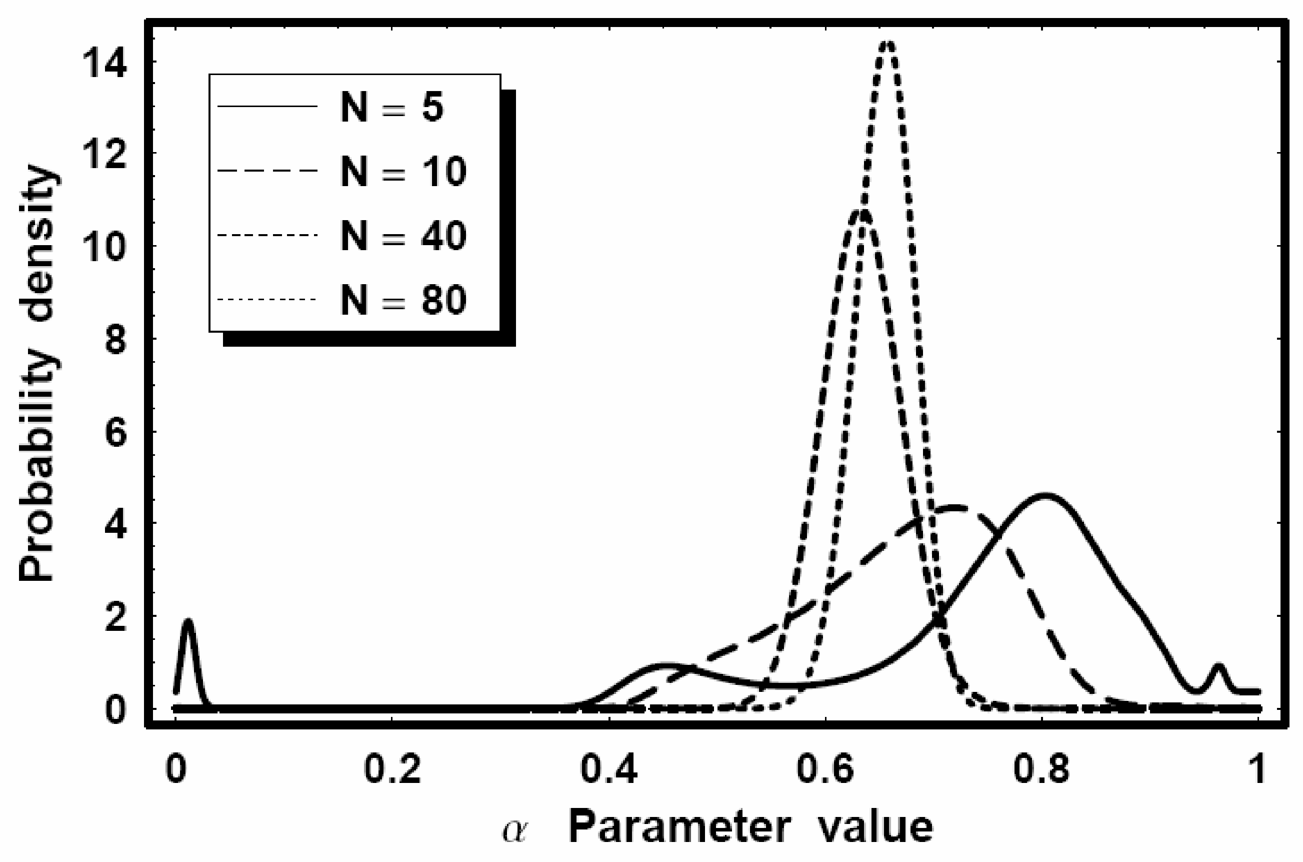
randomized quadrature

adaptive quadrature

High dimensions

MCMC

(chapter 12)



The Bayesian posterior density function for the nonlinear model parameter α for 4 simulated data sets of different size ranging from *N* = 5 to *N* = 80. The *N* = 5 case has the broadest distribution and exhibits 4 maxima.

Mean: Known noise σ

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

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

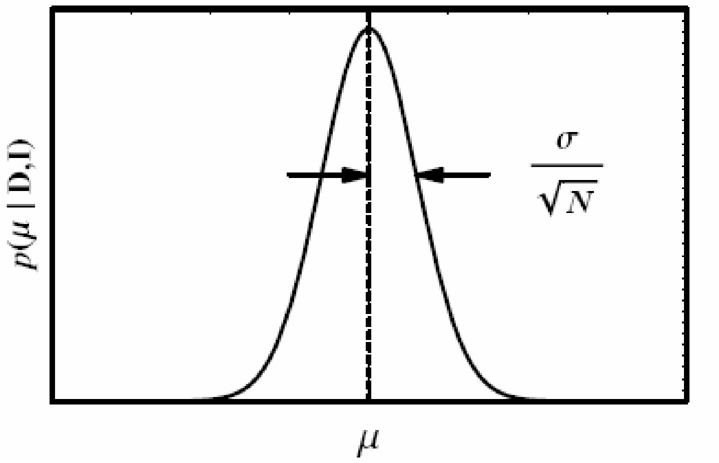
$$\begin{aligned} p(\mu|I) &= K \text{ (constant)} && \mu_{\rm L} \le \mu \le \mu_{\rm H} \\ &= 0 && \text{otherwise.} \end{aligned}$$

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

$$p(\mu|D,I) = \frac{\exp\left\{-\frac{(\mu-\overline{d})^2}{2\sigma^2/N}\right\}}{\int_{\mu_{\rm L}}^{\mu_{\rm H}} \exp\left\{-\frac{(\mu-\overline{d})^2}{2\sigma^2/N}\right\} d\mu}$$



outline

Mean: Known noise, unequal σ

$$
\begin{aligned}
p(D|\mu, I) &= \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(d_i-\mu)^2}{2\sigma_i^2}} \\
&= \left[\prod_{i=1}^{N} \sigma_i^{-1}\right] (2\pi)^{-\frac{N}{2}} \ \exp\left\{-\sum_{i=1}^{N} \frac{(d_i-\mu)^2}{2\sigma_i^2}\right\} \\
&= \left[\prod_{i=1}^{N} \sigma_i^{-1}\right] (2\pi)^{-\frac{N}{2}} \ \exp\left\{-\sum_{i=1}^{N} \frac{w_i(d_i-\mu)^2}{2}\right\}
\end{aligned}
$$

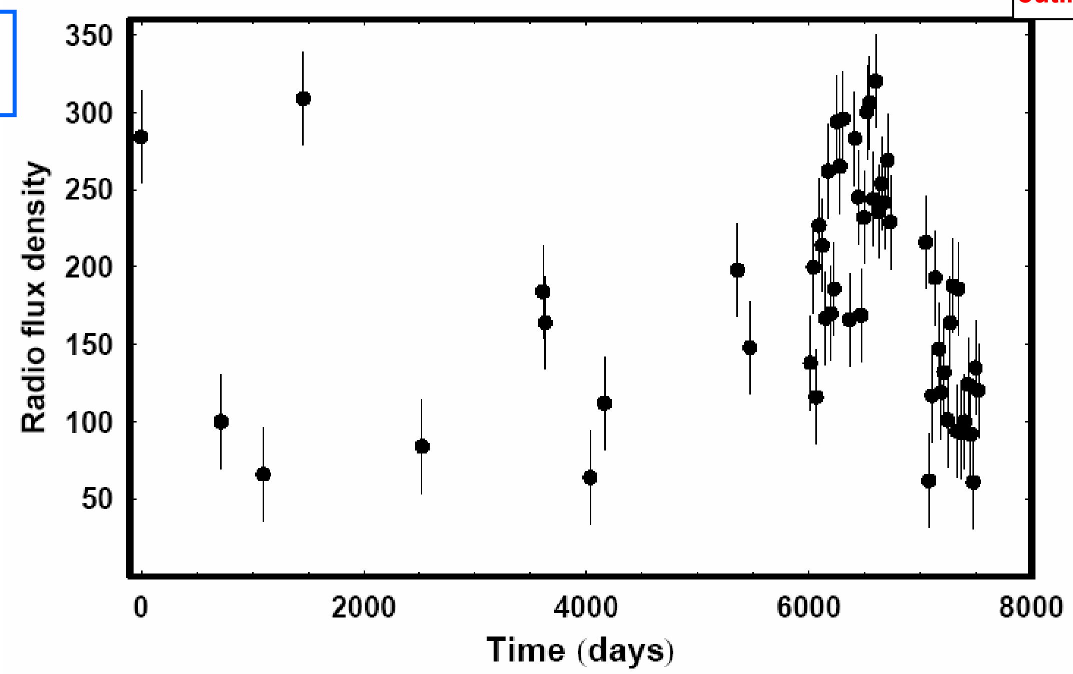
where $w_i = 1/\sigma_i^2$ is called the weight of data value d_i.

Answer

$$
p(\mu|D, I) = \frac{\exp\left\{-\frac{(\mu-\overline{d_w})^2}{2\sigma_w^2}\right\}}{\int_{\mu_{\rm L}}^{\mu_{\rm H}} \ \exp\left\{-\frac{(\mu-\overline{d_w})^2}{2\sigma_w^2}\right\} d\mu}.
$$

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

Mean: Unknown noise σ



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.

outline

Mean: Unknown noise σ

Now we have two unknowns in our model, μ and σ. 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 σ is. In this problem, σ is a nuisance parameter so we marginalize over σ:

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

From the product rule: $p(\mu, \sigma|I) = p(\mu|I)\ p(\sigma|\mu, I)$.
Assuming the prior for σ is independent of the prior for μ, then

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

Mean: Unknown noise σ

outline

We will assume a Jeffreys prior for the scale parameter σ:

$$p(\sigma|I) = \begin{cases} \frac{K}{\sigma} & \sigma_{\rm L} \leq \sigma \leq \sigma_{\rm H} \\ 0 & \text{otherwise.} \end{cases}$$

The constant K is determined from the condition

$$\int_{\sigma_{\rm L}}^{\sigma_{\rm H}} p(\sigma|I) d\sigma = 1 \qquad \Rightarrow \qquad K = \frac{1}{\ln(\frac{\sigma_{\rm H}}{\sigma_{\rm L}})}$$

$$p(\sigma|I) = \frac{1}{\sigma \ln \frac{\sigma_{\rm H}}{\sigma_{\rm L}}}. \tag{9.26}$$

After some maths and a change of variables we can write

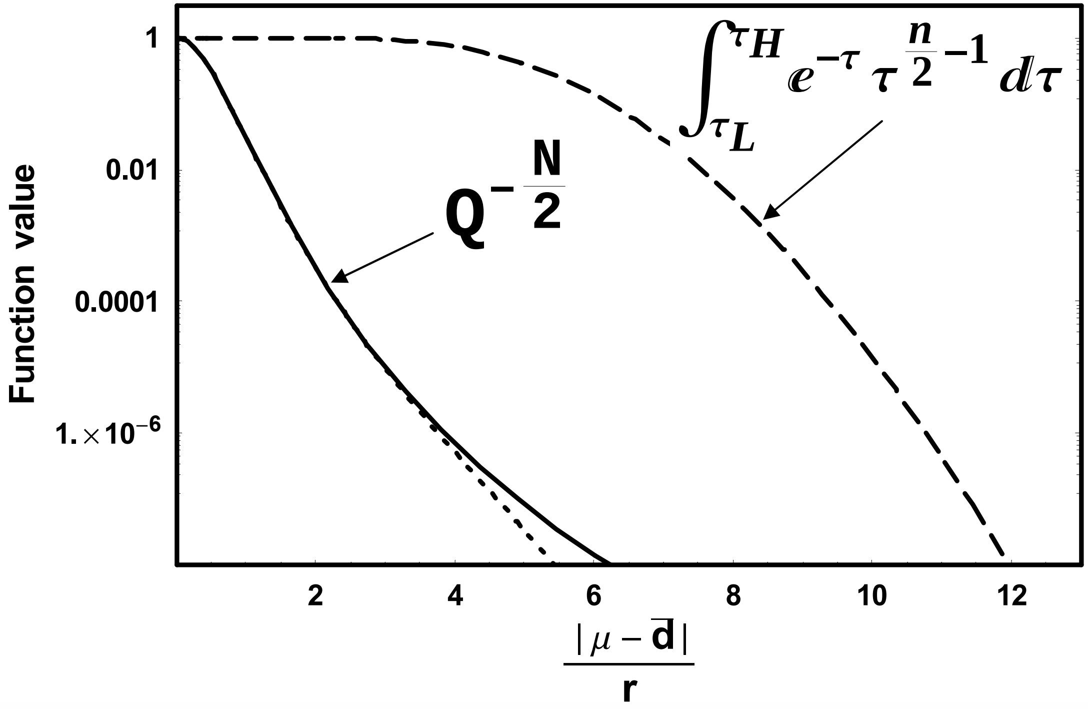
$$\begin{aligned} p(\mu|D,I) &= \frac{Q^{-(\frac{N}{2})} \int_{\tau_{\rm L}}^{\tau_{\rm H}} \tau^{\frac{N}{2}-1} e^{-\tau} d\tau}{\int_{\mu_{\rm L}}^{\mu_{\rm H}} d\mu \, Q^{-(\frac{N}{2})} \int_{\tau_{\rm L}}^{\tau_{\rm H}} \tau^{\frac{N}{2}-1} e^{-\tau} d\tau} \\ &\approx \frac{Q^{-(\frac{N}{2})}}{\int_{\mu_{\rm L}}^{\mu_{\rm H}} d\mu \, Q^{-(\frac{N}{2})}}, \end{aligned} \tag{9.34}$$

where $Q = N(\mu - \overline{d})^2 + Nr^2$**,** $Nr^2 = \sum(d_i - \overline{d})^2$, **and** $\tau = \dfrac{Q}{2\sigma^2}$

Justification for dropping terms involving $\int_{\tau_L}^{\tau_H} \tau^{\frac{N}{2}-1} e^{-\tau} d\tau$

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

N = 10, σ_H = 5 r, σ_L = 0.5 r



$$Q = N\left(\mu - \bar{d}\right)^2 + N\,r^2$$

$$\tau_H = \frac{Q}{2 * \sigma_L^2}$$

$$\tau_L = \frac{Q}{2 * \sigma_H^2}$$

Mean: Unknown noise σ

outline

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

Answer

$$p(\mu|D,I) \approx \frac{\left[1 + \frac{(\mu-\overline{d})^2}{r^2}\right]^{-\frac{N}{2}}}{\int_{\mu_\mathrm{L}}^{\mu_\mathrm{H}} d\mu \, \left[1 + \frac{(\mu-\overline{d})^2}{r^2}\right]^{-\frac{N}{2}}},$$

where the quantity $Nr^2 = \sum(d_i - \overline{d})^2$

Mean: Unknown noise σ

Now compare

$$\left[1+\frac{(\mu-\overline{d})^2}{r^2}\right]^{-\frac{N}{2}} \tag{9.36}$$

with the Student's t distribution

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

If we set

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

and the number of degrees of freedom $\nu = N-1$, then equation (9.36) has the same form as the Student's t distribution .

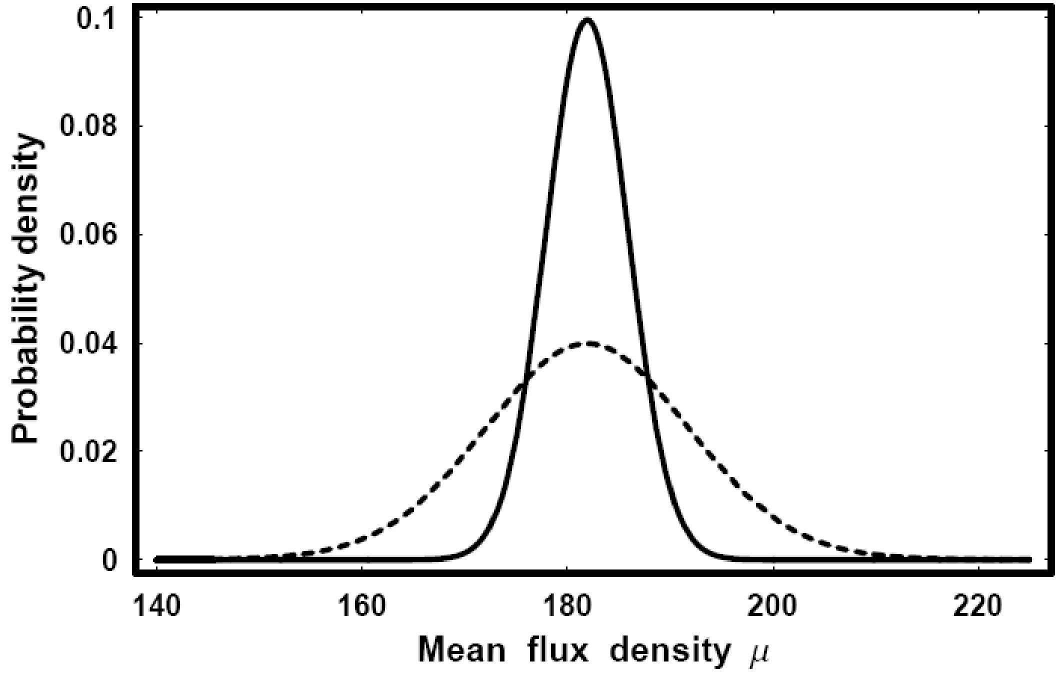


Fig. Comparison of the computed results for the posterior PDF for the mean radio flux density assuming (a) σ known (solid curve), and (b) marginalizing over an unknown σ (dashed curve).

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

This leads to a broader posterior, p(μ | D,I), which reflects the larger effective noise.

outline

Bayesian estimate of σ

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

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.

$$\hat{\sigma} = r$$

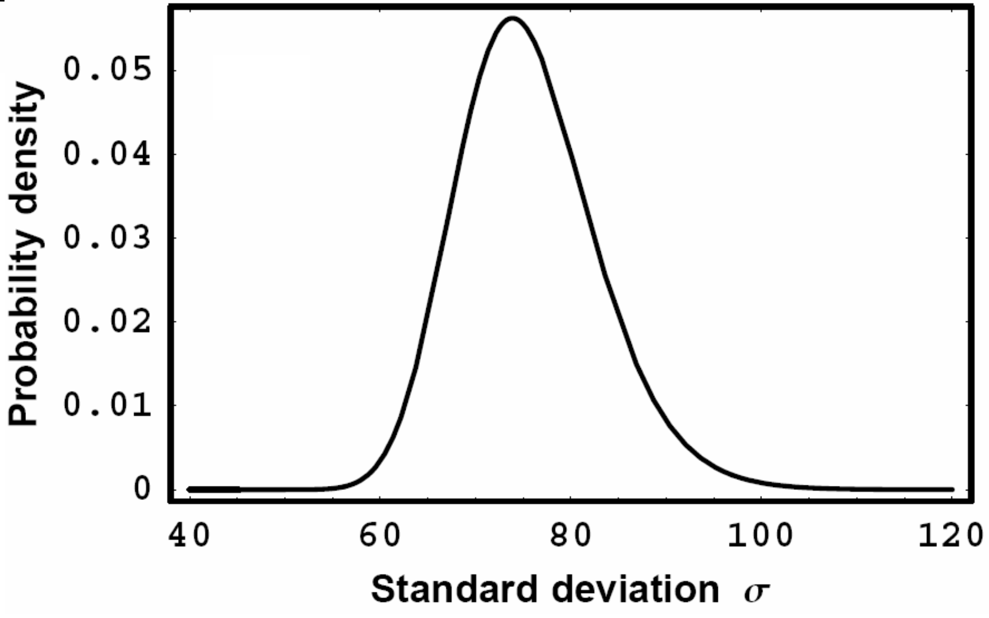
r **= the RMS deviation from** $\overline{d}$.

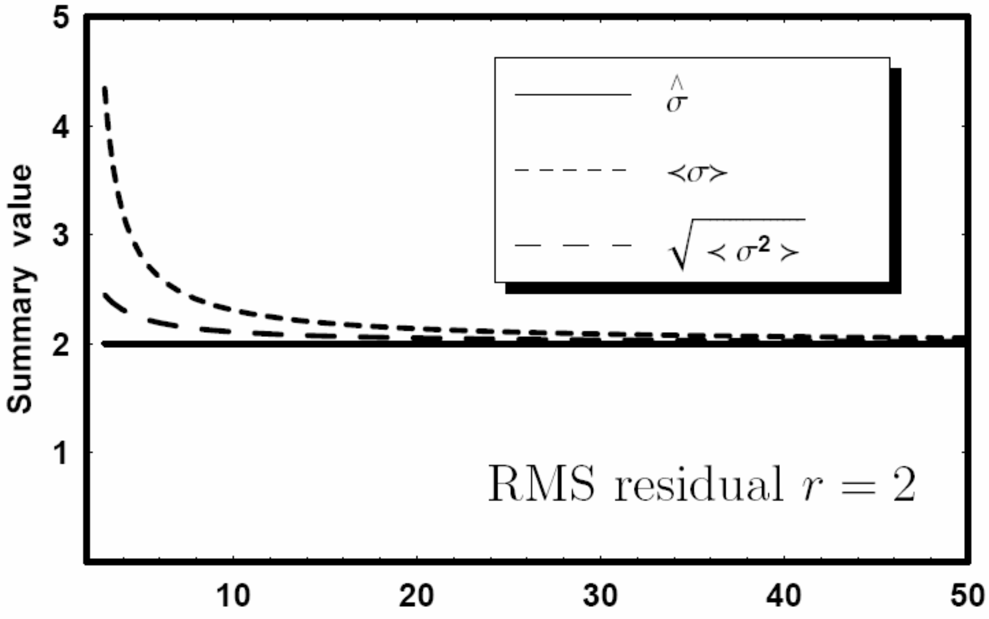
and

$$\langle\sigma^2\rangle = \frac{1}{N-1}\sum_{i=1}^{N}(d_i - \overline{d})^2$$

Compare the later with the frequentist sample variance,

$$S^2 = \sum(d_i - \overline{d})^2/(N-1)$$


Probability density
Standard deviation σ


Summary value
$\hat{\sigma}$
$<\sigma>$
$\sqrt{<\sigma^2>}$
RMS residual $r = 2$