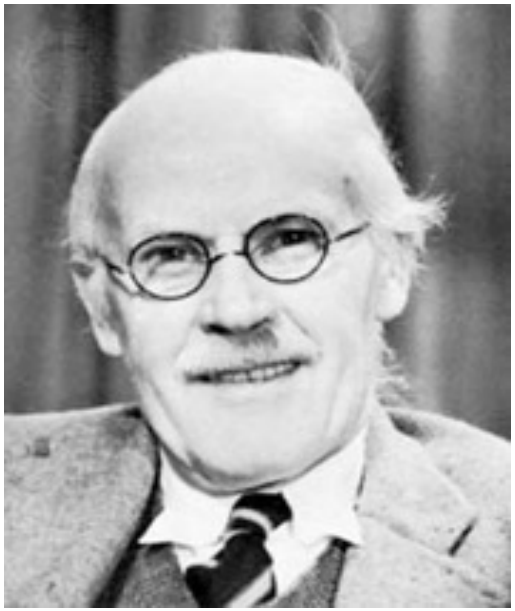
Data modelling - The Bayesian Way



Sir Harold Jeffreys, 1891-1989
Fellow, St John's College, Cambridge 1914-89

Plumian Professor of Astronomy, after many years of teaching maths.
The first to claim a liquid core for the earth.
>400 papers on celestial mechanics, fluid dynamics, meteorology, geophysics, probability plus these books:

The Earth: Its Origin, History and Physical Constitution (1924),
Theory of Probability (1939)
Methods of Mathematical Physics (1946)

'Fisher and Jeffreys first took serious notice of each another in 1933. About all they knew of each other's work was that it was founded on a flawed notion of probability.'

A quick look back at last Tuesday ....

The introduction to data modelling and/or parameter estimation

- We started off with a simple example of **finding a mean for a Gaussian**, knowing that our Gaussian model was correct.

-We then developed a simple Bayesian methodology, showing that the **Bayesian way** was straightforward enough….

-Except for : **expensive modelling** and **difficult integrations**.

-Thus we branched off and looked at the analytical approximations:

Maximimum Likelihood Estimation (MLE)
Least Squares fitting – regression lines being our main example
Minimum chi-square fitting

- We dealt with worked examples for each case.

- This is the time to look back in more detail at the **Bayesian way of modelling**, with its consequent benefits.

Before we leave our friends Maxlik, LS, Chi-Sq ….

Chi-sq: effect of data on fitted parameters is weighted by its variance.
i.e. Large variance => reduced effect on solution.

Consider estimation of the mean, when data on sum have different variances:

We seek e.g. a 'weighted mean'

$$X_w = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_N X_N$$

determining the α_i by requiring minimum variance in X_w. If the X_is are uncorrelated (no systematic errors) the result is

$$X_w = \frac{\frac{X_1}{\sigma_1^2} + \frac{X_2}{\sigma_2^2} + \ldots + \frac{X_N}{\sigma_N^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \ldots + \frac{1}{\sigma_N^2}}.$$

The standard deviation of this weighted mean is

$$\sqrt{\left(\frac{1}{\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \ldots + \frac{1}{\sigma_N^2}\right)}\right)}$$

Our friends Maxlik, LS, Chi-Sq concluded ….

- a particular example of a general technique for constructing linear estimators.

- optimum weight for observation of std dev σ is just $1/\sigma^2$.

- all variances equal => we get the simple average. Squares: premium on accuracy.

- makes the form of chi-sq look plausible.

- these results also follow from ML analysis, assuming Gaussian distributions.

Data Modelling; Parameter Estimation

What are we doing?

Say we have a model. For example,

if our N data Z_i follow a Gaussian distribution

$$\text{prob}(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(z-\mu)^2}{2\sigma^2}\right],$$

then the statistic

$$m = \frac{1}{N}\sum_i Z_i$$

is a good estimator for μ and has a known distribution (a Gaussian again) which can be used for calculating confidence limits.

Or, take the Bayesian way - calculate a probability distribution for μ, given the data.

Any data modelling procedure is just a more elaborate version of this.

Data Modelling: what are we doing? - 2

Suppose our data Z_i were measured at various values of some independent variable X_i, and we believed that they were"really" scattered, with Gaussian errors, around the underlying functional relationship, with $(\alpha_1,\alpha_2, \ldots.)$ unknown parameters (slopes, intercepts, …) of the relationship.

$$\mu = \mu(x,\alpha_1,\alpha_2, \ldots.)$$

We then have

$$\text{prob}(z \mid \alpha_1, \alpha_2 \ldots) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(z-\mu(x,\alpha_1,\alpha_2 \ldots))^2}{2\sigma^2}\right],$$

and, by Bayes' theorem, we have the posterior probability distribution for the parameters

$$\text{prob}(\alpha_1, \alpha_2 \ldots \mid Z_i, \mu) \propto \Pi_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(Z_i-\mu(x,\alpha_1,\alpha_2 \ldots))^2}{2\sigma^2}\right] \text{prob}(\alpha_1, \alpha_2 \ldots)$$

including as usual our prior information. We have included μ as one of the "givens" to emphasize that **everything depends on it being the correct model**.

We are done! We have a probability distribution for the parameters of our model, given the data.

Data Modelling: what are we doing? - 3

Good features:

- it can be used to **update models as new data arrive**, as the posterior from one stage of the experiments becomes the prior for the next.

- we can also deal with **unwanted parameters ("nuisance parameters").**

What are these? Typically we will end up with a probability distribution for various parameters, some of interest (say, cosmological parameters) and some not (say, instrumental calibrations). We can **'marginalize out'** the unwanted parameters by an integration, leaving us with the distribution of the variable of interest that takes account of all plausible values of the unwanted variables.

Data Modelling: what are we doing? - 4

Bad features:

- **Modelling can be very expensive**, always involving finding a maximum or minimum of some merit function. This means evaluating the function, plus its derivatives, many times. The model itself may be the result of a complex computation; evaluating it over a multi-dimensional grid of parameters is even worse.

- **Numerical integration** may be another difficulty. Interesting problems have many parameters. Marginalizing these out, or calculating evidences for discriminating between models, involves multi-dimensional integrals - time-consuming, and hard to check.

  Any analytical help we can get is especially welcome in doing integrations.

  Powerful theorems may allow great simplifications.

Data Modelling: what are we doing? - 5

The most important thing to remember about models is - **they may be wrong.**

Mistaken assumptions about the **distribution of residuals** about the model represent the biggest danger.

- The **wrong parameters** will be deduced from the model.

- **Wrong errors on the parameters** will be obtained.
  It is important to have a range of models, and always to check optimized models against the data.

  Runs Test!

Analytic approximations were developed in past centuries for very good reasons:

- in the limiting case of diffuse priors, the Bayesian approach is very closely related to **maximum likelihood**;
- if the distribution of the residuals from the model is indeed Gaussian, it is closely related to **least squares**.

Bayesian Likelihood Analysis

From Bayes, for model parameters (a vector, in general) $\underline{\boldsymbol{\alpha}}$ and data $\boldsymbol{X_i}$,

$$\mathrm{prob}(\vec{\alpha} \mid X_i) \propto \mathcal{L}(\vec{\alpha} \mid X_i)\mathrm{prob}(\vec{\alpha})$$

However, given **the posterior probability of $\underline{\boldsymbol{\alpha}}$**, we may choose to emphasize properties other than the most probable, e.g. the probability that it is > a certain value.

Two great strengths of the Bayesian approach, plus another one:
(1) ability to deal with **nuisance parameters** via marginalization,
(2) the use of the **evidence or Bayes factor** to choose between models.
(3) the **asymptotic distribution of the likelihood function** itself. $\mathcal{L}(\underline{\boldsymbol{\alpha}})$ is asymptotically a multivariate Gaussian, with **covariance matrix** given by the inverse of the Hessian, evaluated at its peak, namely -

$$\mathbf{F'} = -\begin{bmatrix} \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_1^2} & \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_1 \partial \alpha_2} & \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_1 \partial \alpha_3} & \cdots \\ \frac{\partial \ln \mathcal{L}}{\partial \alpha_2 \partial \alpha_1} & \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_2^2} & \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_2 \partial \alpha_3} & \cdots \\ \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_3 \alpha_1} & \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_3 \partial \alpha_2} & \frac{\partial^2 \ln \mathcal{L}}{\partial \alpha_3^2} & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

- evaluated at the peak, namely the **MLE of $\underline{\boldsymbol{\alpha}}$**

Bayesian Likelihood Analysis, continued ....

The expectation value, or average over many realizations of F', is the

Fisher Information Matrix.

The likelihood then takes the form

$$\mathcal{L}(\vec{\alpha} \mid X_i) = \mathcal{L}(\hat{\vec{\alpha}} \mid X_i) \exp\left(-\frac{1}{2}(\hat{\vec{\alpha}} - \vec{\alpha})^T F (\hat{\vec{\alpha}} - \vec{\alpha})\right)$$

where T denotes the matrix transpose. The integral of this form, over the whole parameter space, is useful:

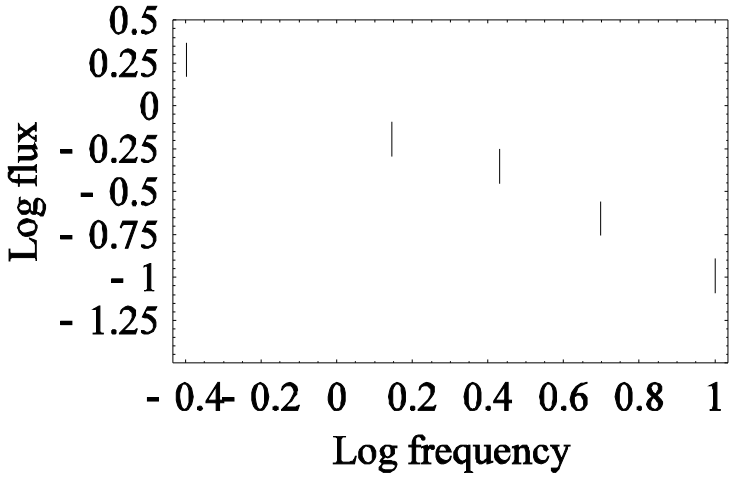
$$\int d\vec{\alpha}\, \mathcal{L}(\vec{\alpha} \mid X_i) = \mathcal{L}(\hat{\vec{\alpha}} \mid X_i) \sqrt{\frac{(2\pi)^k}{|\det F|}}$$

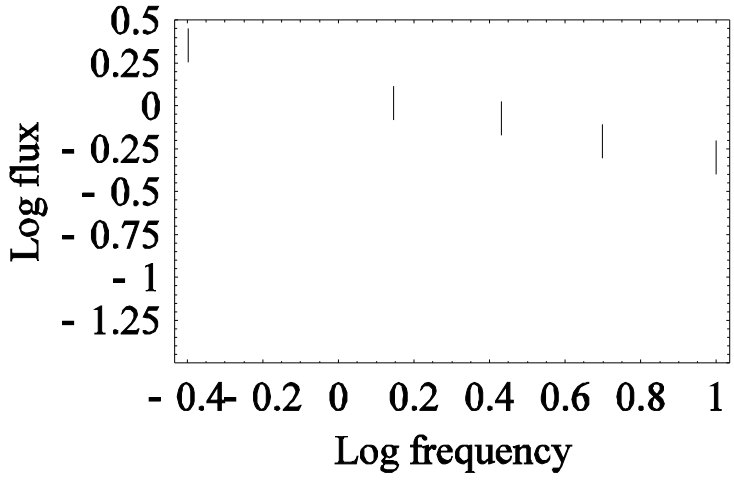
where $\det F$ is the matrix determinant of F and the likelihood is maximized at $\hat{\vec{\alpha}}$. This is called the **Laplace approximation** - see a couple of lectures downstream.

Note here that integrals over *some* of the components of $\vec{\alpha}$ will be potentially more useful – see Jaynes (2003) for the full marginalization integral procedure.

Let us try this approach by developing a **simple two-parameter example**, fitting a power law to some radio flux density data. The example will reappear in different guises, and each time we assume **Gaussian statistics** for the noise, and **uniform (diffuse) priors**.

Bayesian Likelihood Analysis - Example I





Two 'observations' of radio-source spectra. Left: We have noisy flux density measurements at 0.4,1.4, 2.7, 5 and 10 GHz; corresponding data are 1.855, 0.640, 0.444, 0.22 and 0.102 units. Right: same data but with an offset error of 0.4 units as well as random noise.

Call the frequencies f_i and the data S_i. These follow a power law of slope -1, but have a 10% Gaussian noise (= ε) added. Model for f(S) is $k f^{\gamma}$.
Each term in the likelihood product is of the form

$$\frac{1}{\sqrt{2\pi}\epsilon k f_i^{-\gamma}} \exp\left[-\frac{(S_i - k f_i^{-\gamma})^2}{2(\epsilon k f_i^{-\gamma})^2}\right].$$

The likelihood is therefore a function of k and γ.

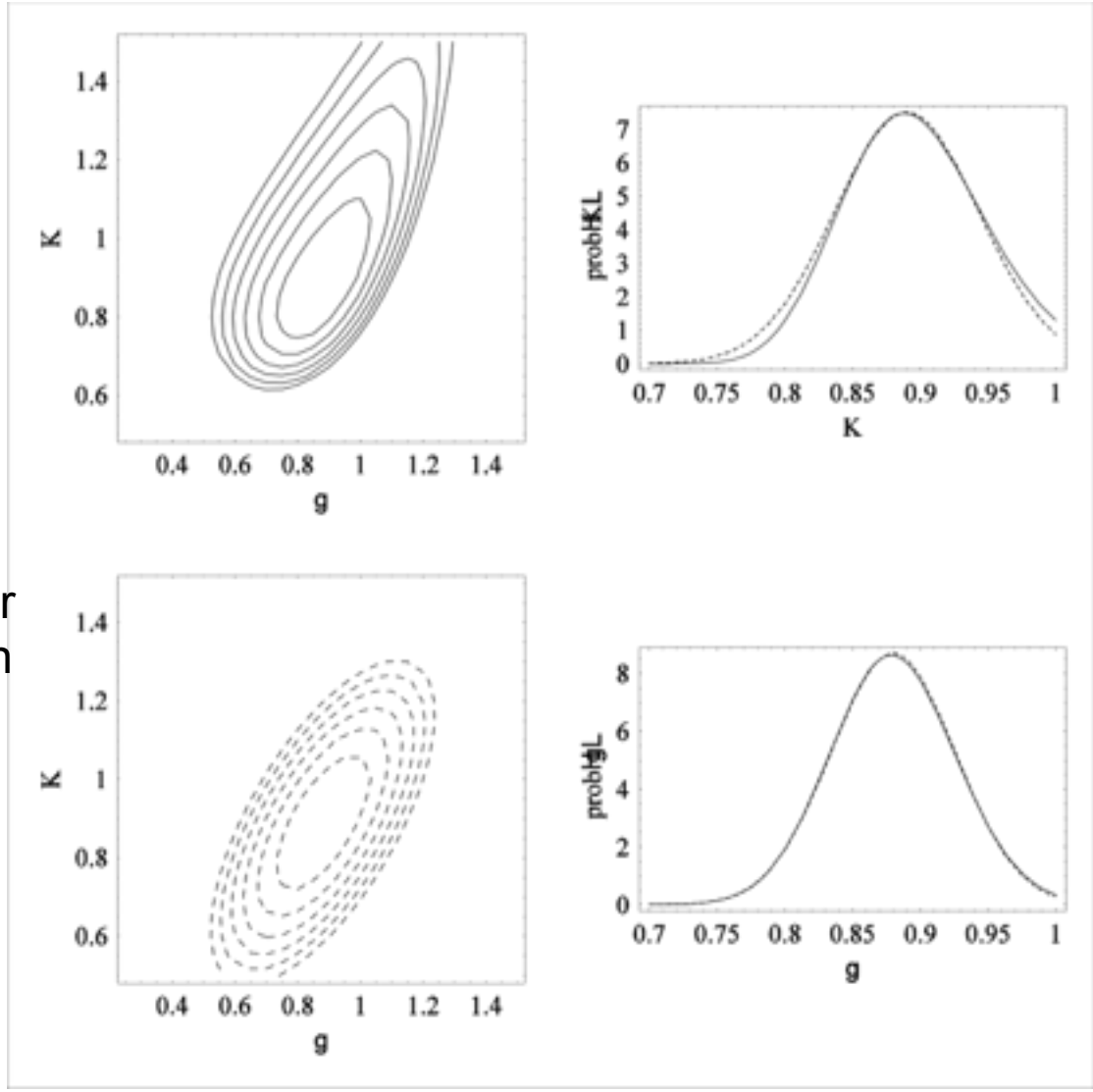
Bayesian Likelihood Analysis - Example 1 cont....

The log likelihood is shown top left.

The calculated Gaussian approximation to the likelihood is bottom left.

There are at least two possibilities for further analysis.

(1) Which pairs of ***(k,γ)*** are, say, 90% probable? Usually a very awkward integration of the posterior probabilities! Multivariate Gaussian approximation to the likelihood is much easier to use: - automatically normalized + there are analytic forms for its integral over any number of its arguments. The areas defined by a particular probability requirement are simple ellipses.



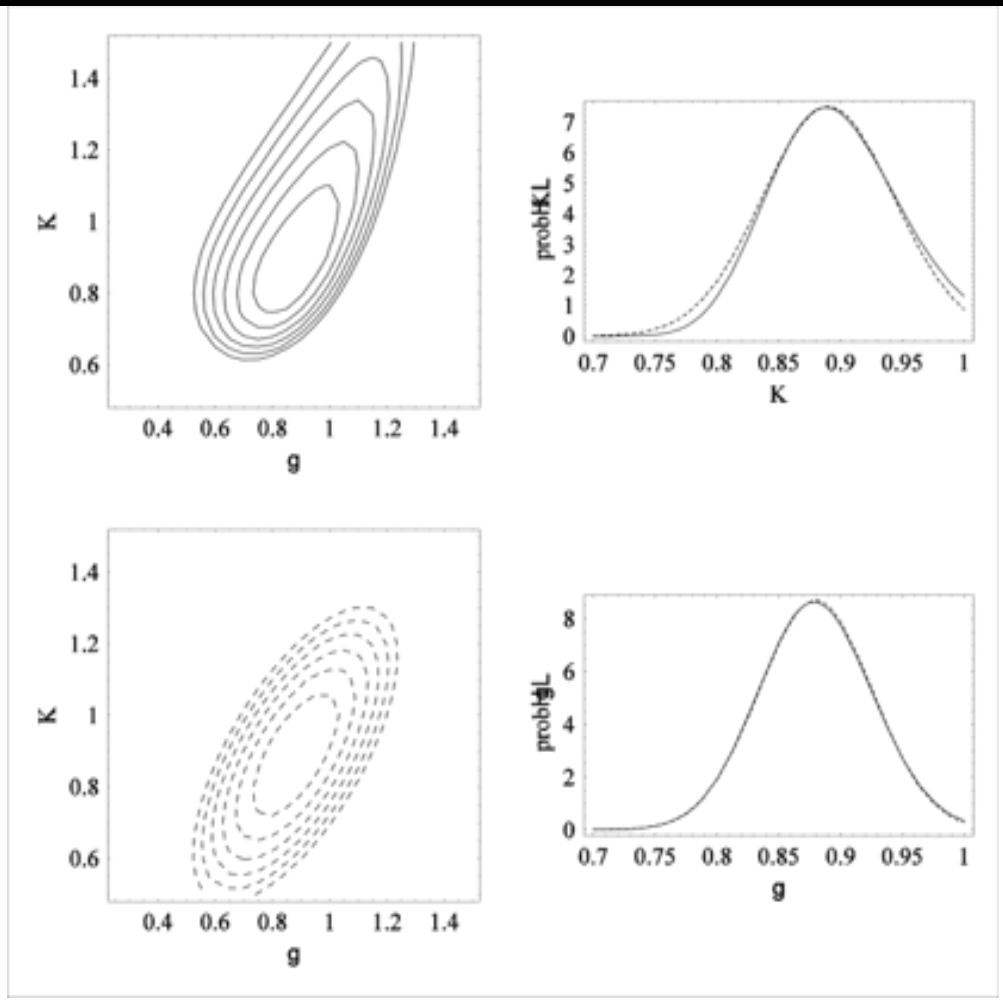
Bayesian Likelihood Analysis - Example 1 done

A second possibility: what is the probability of ***k*** **regardless** of ***γ*** ?

Thus we have the posterior prob(***k***, ***γ*** | S_I) and we form

$$\text{prob}(k \mid S_i) = \int \text{prob}(k, \gamma \mid S_i)\, d\gamma.$$

The probability distributions for ***k*** and ***γ*** are shown as the rightmost set of diagrams in the figure.



Often we are not interested in all the parameters we need to make a model. E.g. If we were investigating radio spectra, we would "**marginalize out**" ***k*** in our example. This is a process of integration over the other parameter(s), in this case just ***k.***

Marginalization

MARGINALIZATION is just integration of the likelihood function (or Bayesian Posterior Distribution) over the useless, non-needed or "nuisance" parameters.

Here, we have rid ourselves of ***k,*** the "height" or intensity of the spectrum, because we have declared ourselves only interested in the "form" of the spectrum as given by the spectral index.

We may also have to estimate **instrumental parameters** as part of our modelling process, but at the end we **marginalize them out** in order to get answers independent of these parameters. **Marginalization will always broaden the distribution of the wanted parameters**, because it is absorbing uncertainty in the parameters we don't want - the **nuisance parameters**. Marginalization is a projection onto an axis. Hence the broadening.

Very useful!

Bayesian Likelihood Analysis - Ex 2, Marginalization

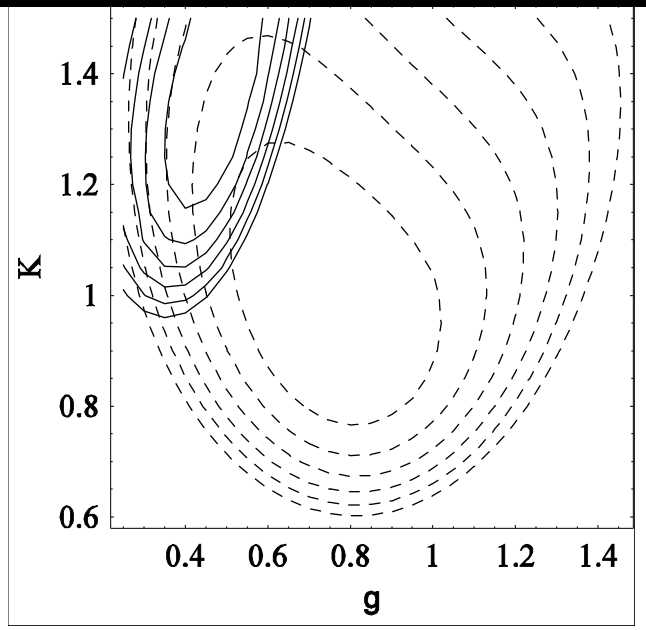
In our radio spectrum example we add (artificially) an offset of 0.4 units to each measurement; the spectrum (earlier Fig) is much flattened as a result.

Calculate two possibilities:

Model A - the simple earlier one, no offsets.

Model B - a model for the flux densities of form $\beta + k f^{\gamma}$. Each likelihood term is then

$$\frac{1}{\sqrt{2\pi}\epsilon k f_i^{-\gamma}} \exp\left[-\frac{(S_i - (\beta + k f_i^{-\gamma}))^2}{2(\epsilon k f_i^{-\gamma})^2}\right].$$



Model A – black contours
Model B – dashed contours

We suspect this offset is present, so we place a prior on β of mean ***0.4***, std dev ε. **Model B** therefore returns a posterior distribution for k, γ and β . **We are not actually interested in β so we marginalize it out**. The likelihoods from the two models are shown : the more complex model does a better job of recovering the true parameters.

Bayesian Likelihood Analysis - Evidence

The procedure works because **there is information in the data about both the instrumental and the source parameters, given the model of the spectrum.** 'Breaks' or 'scale errors' – no chance.

We need to check the "fit" of the two models. In one dimension – been there, done that. In many dimensions, things are harder.

So - **another look at use of evidence (Bayes Factor):**

We choose between **model A** and **model B** - the only possibilities.

prob(α | A) is the prior on ***α*** in model **A**, and similarly for **B**.

The **posterior odds** on model A, compared to model B, are then

$$\frac{\int_\alpha p_A \mathcal{L}(X_i \mid \alpha, A)\mathrm{prob}(\alpha \mid A)}{\int_\alpha p_B \mathcal{L}(X_i \mid \alpha, B)\mathrm{prob}(\alpha \mid B)}$$

and we have to integrate over the range of parameters appropriate to each model. This is worth the effort because we get a straightforward answer to the question:

Which of A or B would it be better to bet on?

Bayesian Likelihood Analysis - Ex 3: the Bet

In the previous two examples we worked out the likelihood functions **L** ***(X$_i$ | k, γ,A)*** for **model A** and similarly for **model B**.

In model B we also have a prior on the offset ***β***:

$$\text{prob}(\beta \mid B) = \frac{1}{\sqrt{2\pi}\,\epsilon} \exp \frac{-(\beta - 0.4)^2}{2(\epsilon)^2}.$$

We form the ratio of the integrals

$$p_A \int dk \int d\gamma \mathcal{L}(X_i \mid k, \gamma, A)$$

and

$$p_B \int dk \int d\gamma \int d\beta \mathcal{L}(X_i \mid k, \gamma, B) \text{prob}(\beta \mid B).$$

Take ***p$_A$ = p$_B$***, an agnostic prior state; and note we have implicitly assumed **uniform priors** on ***k*** and ***γ***.

Cranking through the integrations, we get:

Odds on B compared to A: about 8 to 1.

Bayesian Models of Models

An incorrect model?

Both the deduced parameters and their errors will be wrong.

But circular reasoning often prevails – a) we guess the model and b) try to assess if the deduced parameters are reasonable.

A useful way of expanding the models, as an insurance policy against having the wrong one, is to use **hierarchical models.**

These make use of **hyperparameters.**

In addition to helping with modelling, these notions are useful in the familiar problem of **combining sets of data which have different levels of error.**

BLA - Example 4 : Hierarchical Models

Example 4: Consider example 2 in which we included some kind of offset in the model for each of our flux measurements. Each term in the likelihood function took the form

$$\frac{1}{\sqrt{2\pi}\epsilon k f_i^{-\gamma}} \exp\left[-\frac{(S_i - (\beta + k f_i^{-\gamma}))^2}{2(\epsilon k f_i^{-\gamma})^2}\right].$$

assuming that the offset error ***β*** is the same for each measurement.

Before, we supposed that the distribution of ***β*** was Normal, with a known mean and standard deviation – a strong assumption.

Suppose we knew only the standard deviation, but the mean ***μ*** was unknown. The likelihood is then

$$\exp\left[\frac{-(\beta-\mu)^2}{2\sigma_\beta^2}\right] \prod_i \frac{1}{\sqrt{2\pi}\epsilon k f_i^{-\gamma}} \exp\left[-\frac{(S_i - (\beta + k f_i^{-\gamma}))^2}{2(\epsilon k f_i^{-\gamma})^2}\right]$$

Bayesian Models of Models - Example 4 cont.

μ is now a **hyperparameter**, described by a **hyperprior**. So, for hierarchical models, Bayes' Theorem takes the form

$$\text{prob}(\alpha, \theta \mid X_i) \propto \mathcal{L}(X_i \mid \alpha)\text{prob}(\alpha \mid \theta)\text{prob}(\theta)$$

where $\boldsymbol{X_i}$ are the data and $\boldsymbol{\theta}$ is the hyperparameter (can be a vector). If we integrate out $\boldsymbol{\theta}$, we get a **posterior distribution for the parameter which includes the effect of a range of models.**

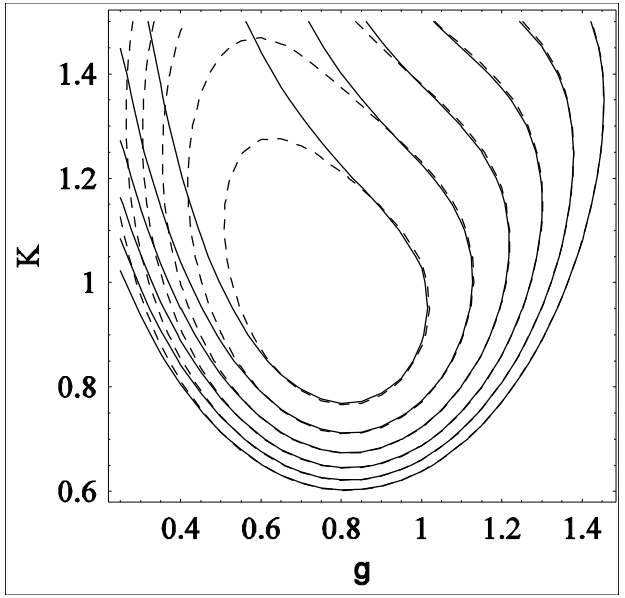
Bayesian Models of Models - Example 4 concluded

In the radio spectrum example, we now take std dev $\sigma_\beta = \varepsilon$ and the prior **prob(μ) = const**.

We compute the likelihood surface by marginalizing over both μ and β.

The integrations are OK because we have Gaussians, and because we integrate from $-\infty$ to $+\infty$. (More realistic integrations, over finite ranges, get very messy.)

In the Fig. we see the likelihood surface for k and γ, compared to the previous "strong" model for which we knew μ. There is a tendency for flatter power laws to be acceptable if we do not \know much about μ.



The log likelihoods for the two models; the black contours are for the hierarchical model and the dashed contours are for known μ .

Bayesian Models of Models, continued

In a more elaborate form of a hierarchical model, we can connect each datum to a separate model, with the models being joined by an overarching structural relationship. In symbols, Bayes then reads

$$\text{prob}(\alpha_i, \theta \mid X_i) \propto \mathcal{L}(X_i \mid \alpha_i)\text{prob}(\alpha_i \mid \theta)\text{prob}(\theta).$$

In a common type of model we may have observations $\boldsymbol{X_i}$ drawn from Gaussians of mean $\boldsymbol{\mu_i}$, with a structural relationship telling us that the $\boldsymbol{\mu_i}$ are in turn drawn from a Gaussian of mean, say, $\boldsymbol{\theta}$.

This is a weaker model than the first sort we considered, because we have allowed many more parameters, linked only by a stochastic relationship. In the case of Gaussians this is a sizeable industry.

Bayesian Models of Models - Example 5

Example 5 - back to our power-law spectrum. If we allow a separate offset $\boldsymbol{\beta}_i$ at each frequency, then each term in the likelihood product takes the form

$$\exp\left[-\frac{(\beta_i-\mu)^2}{2\sigma_\beta^2}\right]\frac{1}{\sqrt{2\pi}\epsilon\, k f_i^{-\gamma}}\exp\left[-\frac{(S_i-(\beta_i+k f_i^{-\gamma}))^2}{2(\epsilon\, k f_i^{-\gamma})^2}\right]$$

and we take again the usual (very weak) prior prob($\boldsymbol{\mu}$) = const.

Marginalizing out each $\boldsymbol{\beta}_i$ by an integration is then exactly the same task for each $\boldsymbol{I}$, and having done this we can compare the likelihood contours with the very first model of these data (no offsets allowed).

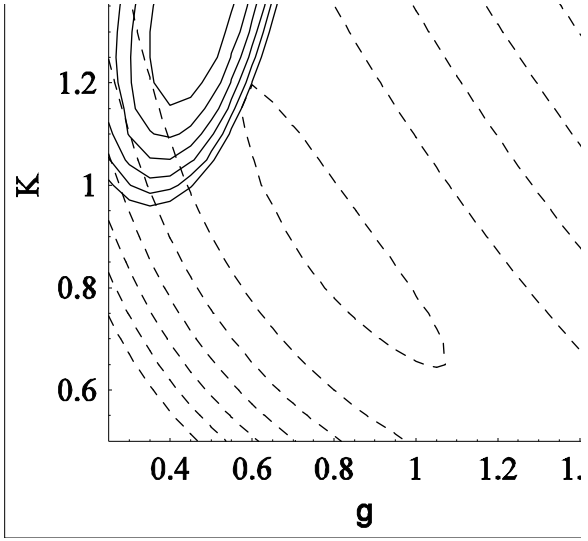
Bayesian Models of Models - Example 5 done

The likelihood contours of the Fig. are instructive!

The hierarchical model, by allowing a range of models, **has moved the solution away from the well-defined (but wrong) parameters of the no-offset model.**

The hierarchical likelihood peaks quite close to the true values of ***(k, y)*** but, but, but …..

Error bounds on these parameters are much wider.



Log likelihoods for the two models. The black contours are for the simplest model, with no provision for offsets. Dashed contours are for the weak hierarchical model, allowing separate offsets at each frequency.

Bayesian Models of Models - Example 5 done

This is a general message:

Allowing uncertainty in our models may make the answers appear less precise, but is an **insurance against well-defined but wrong answers from modelling**.

Finally, note **broadening the range of models** is a useful technique in **combining data.**

The idea is to allocate weights ξ_x and ξ_y to two datasets in which the 'normal' weights (e.g $1/\sigma^2$) appear not to serve. (I.e. the data disagree on the basis of our error estimates BUT we still want to use them all.)

This is a hierarchical model, and the weights are hyperparameters (Hobson, Bridle and Lahav 2002). Details and an example follow pp 171-176 of Practical Statistics, 2nd ed.