ASTR509 - 12

Model choice

William of Ockham (or Occam)
Philosopher, rebel priest ~1285 to 1349

Entered Franciscan order, studied theology at Oxford,

Bad ideas => degree witheld.

Summoned to the Papal Court at Avignon 1324, house arrest, branded as heretic, excommunicated. Escaped in 1328; then chased all over Europe by irate Popes.

His bad ideas continued to develop – 'apostolic poverty', the evils of Papal power, the need for civil sovereignty, the rule of mediaeval parsimony (principle of economy) => Ockham's Razor

Died Munich 1349 in/before the Black Death plague

'Frustra fit per plura quod potest fieri per pauciora' – `It is futile to do with more things that which can be done with fewer ', or 'plurality should not be assumed without necessity', or in modern English, 'keep it simple, stupid')

Remember where we were?

The introduction to data modelling and/or parameter estimation

- We reviewed the core of Bayesian data modelling and parameter estimation.

- Good features (1) updating, (2) marginalization, (3) asymptotic distribution of the Likelihood Function (Fisher Matrix etc, more later)

- Bad features: (1) expensive (but who cares nowadays), (2) tough integrations (but we will look at how to deal with this)

- We spent the rest of the time on a two-parameter example, our 5-frequency spectral measurement of a mythical extragalactic radio source.

- This example showed how we could get rid of an unwanted instrumental effect, the addition of a constant to each flux at each of the 5 frequencies – the process of marginalization. We showed that this worked in the instance of having some prior information (a) presence, (b) constancy at each freq, and (c) some guess of its magnitude.

- We had a first look at Evidence, and we worked out the odds of whether the naïve model or the model with the offset better described the 'damaged' data. The latter wins by odds of 8:1.

- We looked at further model expansion to guard against assumption of too naïve a model: hierarchical models, hyperparameters – useful for combining data sets giving disparate answers.

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

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

where ***X*** $\boldsymbol{X_i}$ 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.**

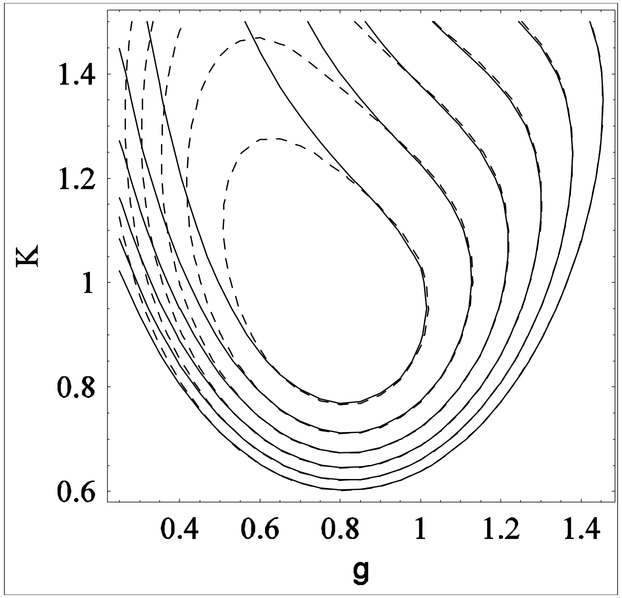
Bayesian Models of Models - Example 4 concluded

In the radio spectrum example, we now take std dev $\boldsymbol{\sigma_\beta}$ = $\boldsymbol{\varepsilon}$ 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 -∞ to +∞. (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

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

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 ***l***, and having done this we can compare the likelihood contours with the very first model of these data (no offsets allowed).

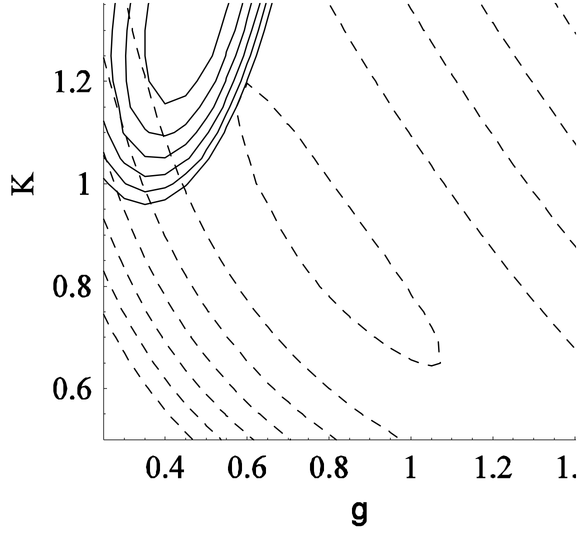
Bayesian Models of Models - Example 5 done

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, γ)*** 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 $\boldsymbol{\xi_x}$ and $\boldsymbol{\xi_y}$ to two datasets in which the 'normal' weights (e.g $\boldsymbol{1/\sigma^2}$) 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.

Data Modelling: Choosing the Model

The Bayesian way provides a principled scheme for making choices between models.

Classical testing: say fit a model via LS, and then use X^2 to decide if we should reject it. But what if

- the deviations are likely?
- several models are available?

NB if a model is correct, the significance level from a X^2 test (or any significance test!) will be uniformly distributed between 0 and 1.

For model H and data D, a significance level from min X^2 tells us about the

conditional probability prob(D|H)

But we want

conditional prob (H|D)

We can't get this from prob(D|H) without an application of Bayes' Theorem and its involvement with prior probabilities.

Model Choice and Bayesian Evidence

Suppose we have just two models H_1 and H_2, with parameter sets α and β.

Set of data D. Then Bayes' Theorem for the posterior probs:

$$\mathbf{prob}(H_1, \vec{\alpha} \mid D) = \frac{\mathbf{prob}(D \mid H_1, \vec{\alpha})\, \mathbf{prob}(\vec{\alpha} \mid H_1)\mathbf{prob}(H_1)}{E}$$

$$\mathbf{prob}(H_2, \vec{\beta} \mid D) = \frac{\mathbf{prob}(D \mid H_2, \vec{\beta})\, \mathbf{prob}(\vec{\beta} \mid H_2)\mathbf{prob}(H_2)}{E}.$$

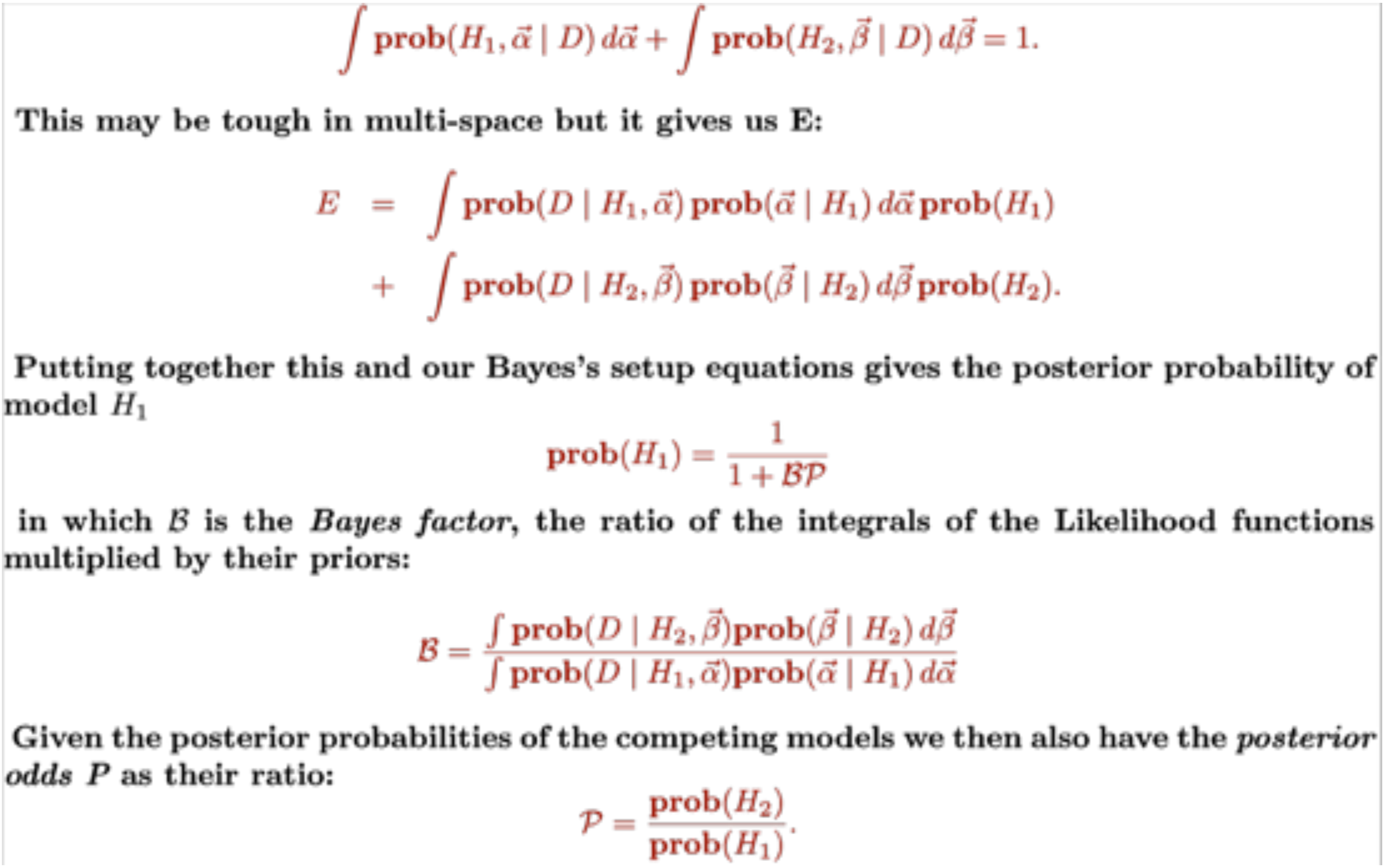
Note we've doubled up on priors!

- priors on the models – our degree of belief that we've got it right
- priors on parameters – here we put in our known contraints or beliefs

E is normalizing factor to make LHS a probability – its importance is coming…..

Model Choice and Bayesian Evidence 2

We can find E :



$$\int \mathbf{prob}(H_1, \vec{\alpha} \mid D)\, d\vec{\alpha} + \int \mathbf{prob}(H_2, \vec{\beta} \mid D)\, d\vec{\beta} = 1.$$

This may be tough in multi-space but it gives us E:

$$\begin{aligned} E &= \int \mathbf{prob}(D \mid H_1, \vec{\alpha})\, \mathbf{prob}(\vec{\alpha} \mid H_1)\, d\vec{\alpha}\, \mathbf{prob}(H_1) \\ &+ \int \mathbf{prob}(D \mid H_2, \vec{\beta})\, \mathbf{prob}(\vec{\beta} \mid H_2)\, d\vec{\beta}\, \mathbf{prob}(H_2). \end{aligned}$$

Putting together this and our Bayes's setup equations gives the posterior probability of model H_1

$$\mathbf{prob}(H_1) = \frac{1}{1 + \mathcal{BP}}$$ 1

in which $\mathcal{B}$ is the *Bayes factor*, the ratio of the integrals of the Likelihood functions multiplied by their priors:

$$\mathcal{B} = \frac{\int \mathbf{prob}(D \mid H_2, \vec{\beta})\mathbf{prob}(\vec{\beta} \mid H_2)\, d\vec{\beta}}{\int \mathbf{prob}(D \mid H_1, \vec{\alpha})\mathbf{prob}(\vec{\alpha} \mid H_1)\, d\vec{\alpha}}$$ 2

Given the posterior probabilities of the competing models we then also have the *posterior odds* P as their ratio:

$$\mathcal{P} = \frac{\mathbf{prob}(H_2)}{\mathbf{prob}(H_1)}.$$ 3

Model Choice and Bayesian Evidence 3

- last three equations encapsulate the Bayesian model choice method

- key ingredient – **BAYES FACTOR**, a ratio of the terms sometimes called **EVIDENCE**

- **EVIDENCE** terms are the average of the Likelihood Function over the Prior on the parameters

- relative magnitude of the **EVIDENCE** for each model determines its posterior probability

- normalizing term E is sum of **EVIDENCE** terms, each weighted by Prior on relevant model

Model Choice and Bayesian Evidence 4

Simple example:

Prior odds **P** are large – i.e. from previous experience, H_2 is probably correct

But data are much more likely under H_1, so we have a small Bayes factor **B**

If Bayes factor is small enough it will outweigh the large prior odds, so that

prob(H_1) = 1/(1 + **BP**) ≈ 1

Thus the data have modified the conclusion:

H_1 is now probably correct

Model Simplicity and the 'Ockham Factor'

NB: parameter priors are within the evidence integrals. Significance?

Consider case of narrow likelihood functions, and single-parameter models, α for H_1 and β for H_2, with ranges Δα and Δβ, so that prob(α)=1/Δα and prob(β)=1/Δβ. This gives

BP = (ratio of likelihood integrals) x (Δα/ Δβ) x (ratio of prior odds)

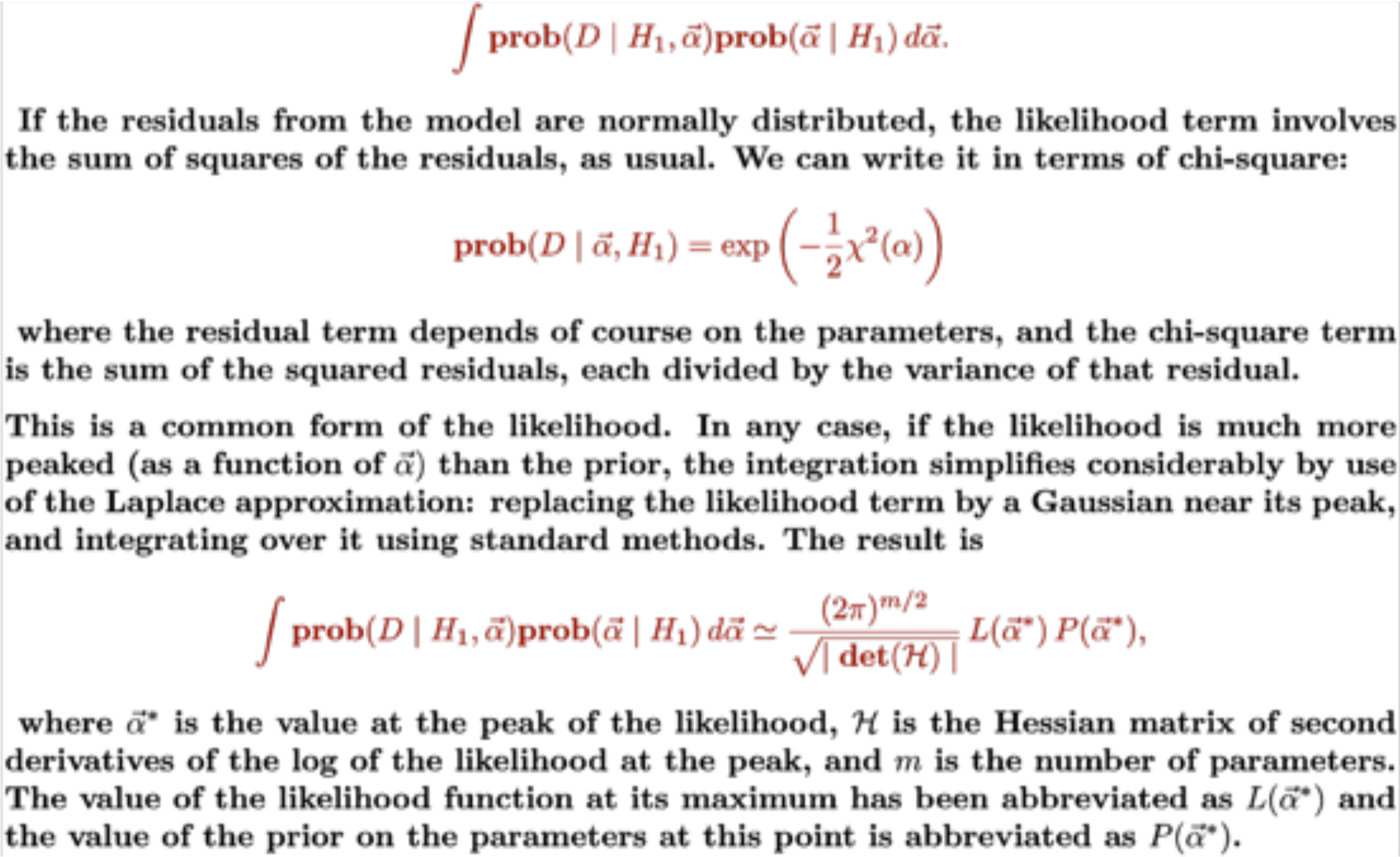
In our previous example, the prior odds ratio was large (H_2 favoured). Suppose our prior about its parameter β is vague (broad) compared to the H_1 parameter α, i.e. Δα/ Δβ << 1. Then the data have to work very strongly in favour of H_2 for it to come out with the larger posterior probability, much harder than if Δα ≈ Δβ.

This is the so-called Ockham factor at work.

The model H_1 is 'simpler' than H_2 because it is more specific, less prior range, and this 'simplicity' boosts its posterior probability. It's a built in mechanism – there's no real factor involved – we only got one by the above shortcut.

Avoiding the Integrations

The heart of the matter is the integration of a LikeFn over its params and prior:



$$\int \mathbf{prob}(D \mid H_1, \vec{\alpha})\mathbf{prob}(\vec{\alpha} \mid H_1)\, d\vec{\alpha}.$$

If the residuals from the model are normally distributed, the likelihood term involves the sum of squares of the residuals, as usual. We can write it in terms of chi-square:

$$\mathbf{prob}(D \mid \vec{\alpha}, H_1) = \exp\left(-\frac{1}{2}\chi^2(\alpha)\right)$$

where the residual term depends of course on the parameters, and the chi-square term is the sum of the squared residuals, each divided by the variance of that residual.

This is a common form of the likelihood. In any case, if the likelihood is much more peaked (as a function of $\vec{\alpha}$) than the prior, the integration simplifies considerably by use of the Laplace approximation: replacing the likelihood term by a Gaussian near its peak, and integrating over it using standard methods. The result is

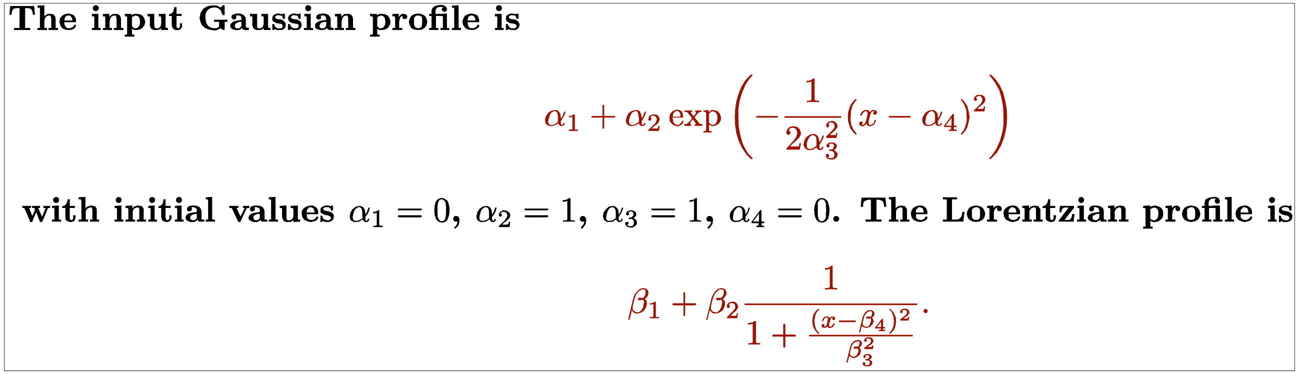
$$\int \mathbf{prob}(D \mid H_1, \vec{\alpha})\mathbf{prob}(\vec{\alpha} \mid H_1)\, d\vec{\alpha} \simeq \frac{(2\pi)^{m/2}}{\sqrt{\mid \mathbf{det}(\mathcal{H}) \mid}} L(\vec{\alpha}^*)\, P(\vec{\alpha}^*),$$

where $\vec{\alpha}^*$ is the value at the peak of the likelihood, $\mathcal{H}$ is the Hessian matrix of second derivatives of the log of the likelihood at the peak, and m is the number of parameters. The value of the likelihood function at its maximum has been abbreviated as $L(\vec{\alpha}^*)$ and the value of the prior on the parameters at this point is abbreviated as $P(\vec{\alpha}^*)$.

So: find the max posterior prob, evaluate **H** and determinant. Works well on smallish problems. Must keep track of normalizing factor of priors - brings Ockham factor into play.

Let’s try it ....

We are trying to decide if a single spectral line is Gaussian or Lorentzian. Parameters in each case: baseline level, height, width, centre. Gaussian model H_1, Lorentzian H_2.



The input Gaussian profile is

$$\alpha_1 + \alpha_2 \exp\left(-\frac{1}{2\alpha_3^2}(x-\alpha_4)^2\right)$$

with initial values $\alpha_1 = 0$**,** $\alpha_2 = 1$**,** $\alpha_3 = 1$**,** $\alpha_4 = 0$**. The Lorentzian profile is**

$$\beta_1 + \beta_2 \frac{1}{1+\frac{(x-\beta_4)^2}{\beta_3^2}}.$$

Generate data from the Gaussian model, adding Gaussian random noise

Then fit Gaussian and Lorentzian profiles, using the known noise std deviation and using Least Squares => sums of squared residuals, normalized.

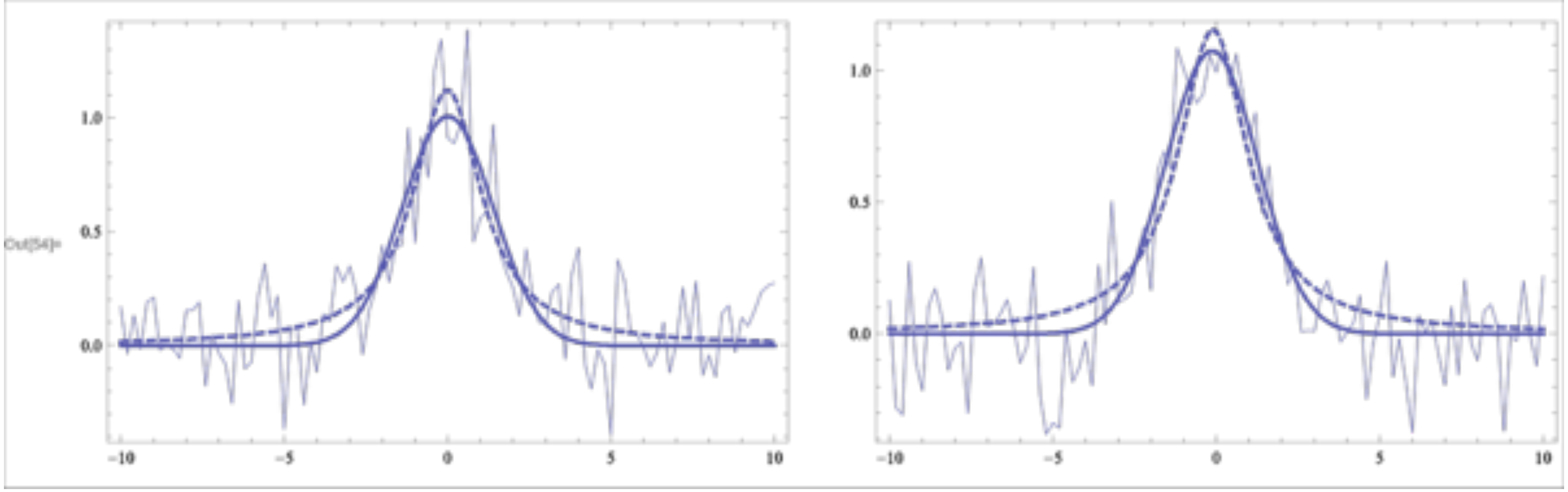
Assume flat priors, same ranges for all params. Assume each model equally likely.

Run MC simulations, and

Compute Bayes Factor by Laplace Approximation.

What do we get?

We are trying to decide if a single spectral line is Lorentzian or Gaussian. Four parameters in each case: baseline level, height, width, centre. Gaussian model is H_1, Lorentzian is H_2.

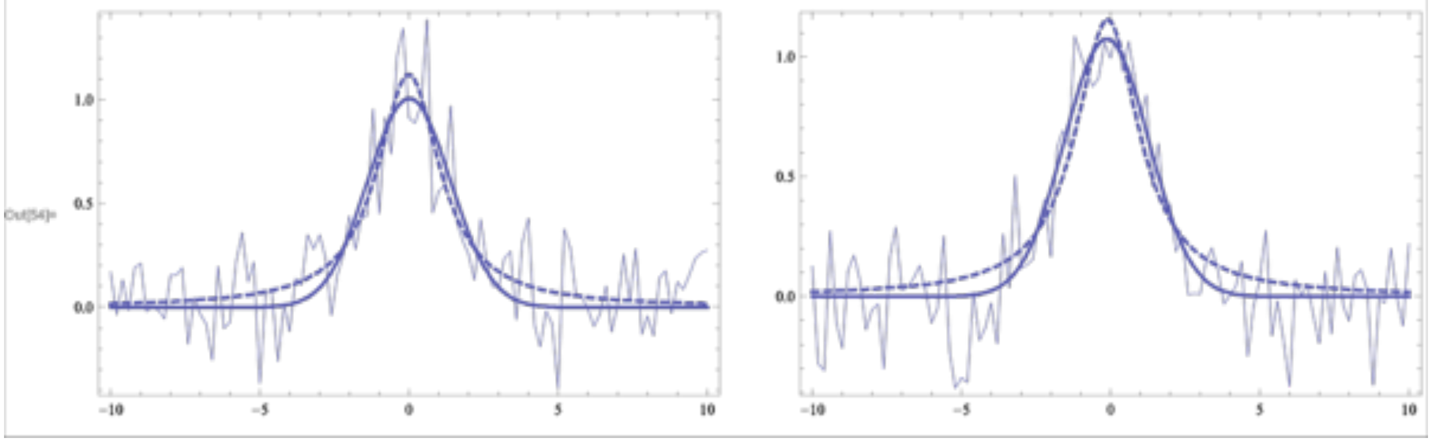


Signal-to-noise ratio at peak of line = 5, Lorentzian simulation in left panel, Gaussian in right. Lorentzian fit favoured in left panel, odds 30:1. Gaussian fit favoured in right panel, odds 100:1.

What?? What do you mean it's OK? The eye says these conclusions are probably in the right senses - but this level of certainty? No way.

BF is a statistic : the odds show much variation.

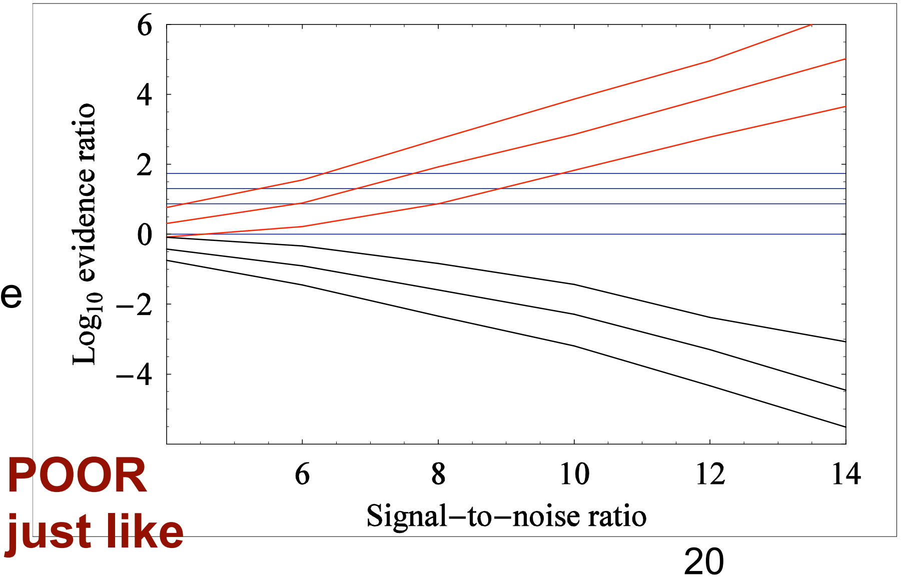
Bayes Factor is a Statistic ….



Signal-to-noise ratio at peak of line = 5. Lorentzian fit favoured in left panel, odds 30:1. Gaussian fit favoured in right panel, odds 100:1.

BF is a ***statistic : the odds show much variation. And note size of odds!***

Run the exp’t many times: the picture on the left emerges. Red=Gaussian, upper/ lower lines are 25 and 75 percentiles. Black=Lorentzian, Central line=median, Blue lines mark odds on Gaussian that are even, e^2:1, e^3:1,e^4:1



Ability to pick Gaussian when true is POOR at low s/n. Just like classical testing; just like real life. Need goodness-of-fit test.

Other Methods of Model Choice

The 'Akaike' and 'Bayesian' Information Criteria:

Two other commonly used criteria: AIC and BIC, defined as:

$$\mathrm{AIC} = -2\log\mathcal{L}_{\mathbf{max}} + 2k$$

and

$$\mathrm{BIC} = -2\log\mathcal{L}_{\mathbf{max}} + k\log N$$

- First term is maxlik; second is penalty term, k params, N data points

- Only relative values used; pick model with smallest AIC or BIC

- Easier to use than Bayesian analysis with integration to get Evidence

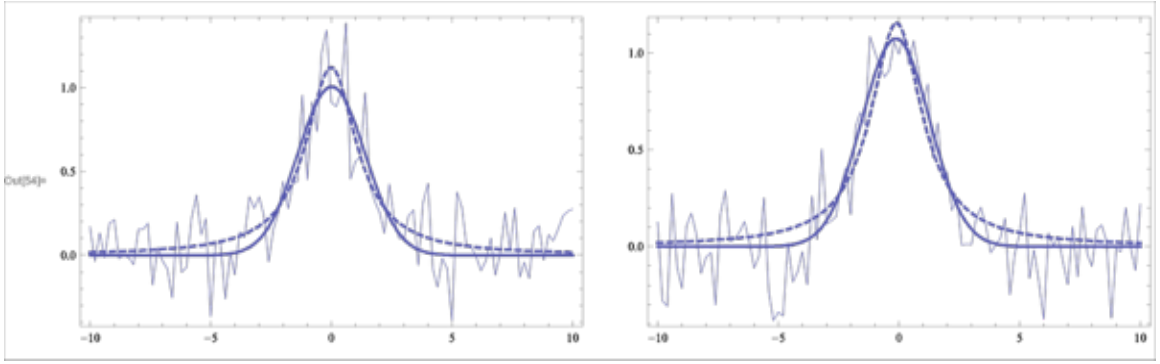
- BIC is an approx to full Evidence

- With same no. of params, diffuse priors, approxs are as good as full Bayes

- with different param numbers, differences emerge with even simple priors

- large N + Gaussian errors => minimum X^2

Akaike and Bayesian IC: Example



Add some curvature to baseline of Lorentzian model; offset its wider wings to make it more competitive with Gaussian model.

Assume same priors on line-shape params

Assume baseline term is $(\beta_5 x^2 + \beta_6 x^4)$, flat priors of widths $\Delta\beta_5$, $\Delta\beta_6$

Rough prior: max in both terms < line height, ie $\Delta\beta_5=10^{-2}$ and $\Delta\beta_6=10^{-4}$

The detailed sum shows that the odds on the Gaussian get shortened by

about 3 orders of magnitude - **We are close to evens now!**

The unequal number of parameters highlights how the Bayes Factor approach differs fundamentally from minimum X^2.

Now the priors really matter.