8 How to compare models

8.1 General approach

In some cases it would be nice to know the probability that a model M used in the analysis of data $\mathbf{y}$ is indeed the correct model. This means that we would like to know the probability

$$\mathrm{prob}(M|\mathbf{y}, I).$$

The problem, however, is here, how to normalize this distribution. We would have to sum over all possible models, but we do not know them! A way to avoid the normalization problem is, to compare such probabilities by taking ratios. This procedure leads to the so-called *significance tests*.

Suppose we have two different physically motivated models M_1 and M_2 at hand. We would like to know, which of the two models is more strongly supported by the measured data $\mathbf{y}$. Then we may look at the probability ratio

$$\frac{\mathrm{prob}(M_1|\mathbf{y}, I)}{\mathrm{prob}(M_2|\mathbf{y}, I)}.$$

This quantity is still quite asymmetric regarding the two models. We would prefer M_1 over M_2, if this ratio takes values between 1 and ∞, but we would prefer M_2 over M_1, if the ratio is between zero and one. A more symmetric measure is the logarithm

$$\ln \frac{\mathrm{prob}(M_1|\mathbf{y}, I)}{\mathrm{prob}(M_2|\mathbf{y}, I)}$$

of the above ratio. In this case we would prefer M_1 over M_2, if this quantity is larger than zero, and we would prefer M_2 over M_1, if it is smaller than zero. Note that prefering one model over the other does not at all mean to *reject* the other model!

Using Bayes' theorem in the form

$$\mathrm{prob}(M_i|\mathbf{y}, I) = \frac{\mathrm{prob}(M_i|I)\mathrm{prob}(\mathbf{y}|M_i, I)}{\mathrm{prob}(\mathbf{y}|I)}. \tag{52}$$

we can rewrite the expression giving

$$\ln \frac{\mathrm{prob}(M_1|\mathbf{y}, I)}{\mathrm{prob}(M_2|\mathbf{y}, I)} = \ln \frac{\mathrm{prob}(M_1|I)}{\mathrm{prob}(M_2|I)} + \ln \frac{\mathrm{prob}(\mathbf{y}|M_1, I)}{\mathrm{prob}(\mathbf{y}|M_2, I)}. \tag{53}$$

While the first term on the right hand side describes our preference for M_1 and M_2 before we know the data $\mathbf{y}$, the second term is the logarithm of the ratio of the evidences for $\mathbf{y}$

given the two alternative models. The ratio of evidences is called the *Bayes factor*. The first term is only of relevance, if we have any prior information (e.g., a previous measurement) that would make us prefer one model over the other. If this is not the case, we would give equal prior probabilities to the two models, and the logarithm of their ratio would therefore be zero. Given no prior preference, we would therefore prefer the model that is more likely having produced the data $\mathbf{y}$. Figure 51 visualizes the basic model comparison rule in eq. (51) under the assumption that we have no prior preference for any of the two models. In general, a more complex, or more flexible model, like model 2 in the figure, will spread its probability distribution over a larger range in data space, because it is able to generate a larger range of data, but it must be normalized to 1. In contrast a simpler more stringent model cannot produce data in such a large range, but as a consequence, the probability density is higher in that region, compared to model 2. Depending on where on the data axis the specific data is that we have obtained, we will prefer the model with the higher probability density in this location.

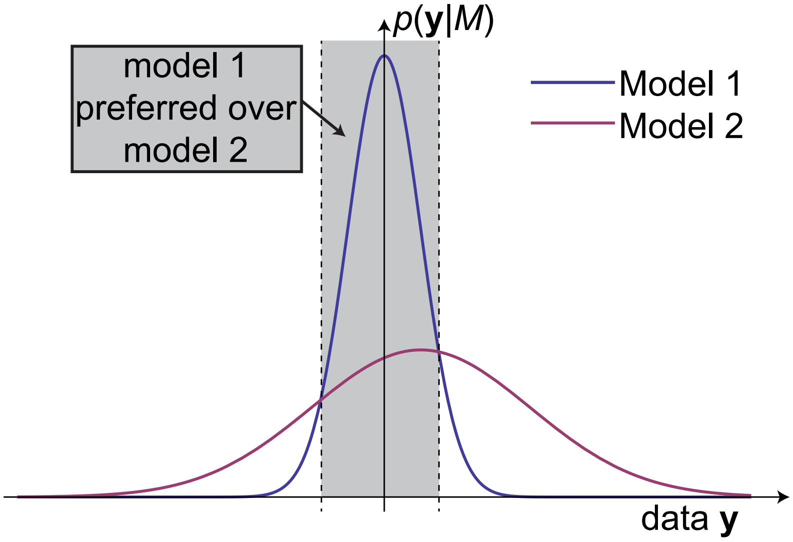

Abbildung 51: Schematic visualization of the model comparison rule in eq. (modelsel), assuming that the prior probabilities of the two models are identical. Model 2 is preferred over model 1 in all regions of data space, except for the gray-shaded region, where model 1 is preferred over model 2.

The evidences are given by

$$\mathrm{prob}(\mathbf{y}|M_i, I) = \int d\boldsymbol{\theta}_i \mathrm{prob}(\boldsymbol{\theta}_i|M_i, I)\mathrm{prob}(\mathbf{y}|\boldsymbol{\theta}_i, M_i, I), \tag{54}$$

where $\boldsymbol{\theta}_i$ represents the set of parameters of model M_i. We see here that the prior probability influences the distribution of the evidence. In this sense it is an integral part of the model. For example, we may consider two models that share a common likelihood, but differ in their prior probability distributions.

First Example. Consider the measurement of the length of an object using a ruler. Suppose we have prior knowledge about the length of the ruler given by a gaussian distribution centered around L_0 and width σ_L. The likelihood for measuring a particular value ℓ is also taken to be a normal distribution centered around the true value with reading error σ, the precision of the measurement. For this case, the evidence can be worked out to be

$$\mathrm{prob}(\ell|L_0, \sigma_L, \sigma, M, I) = \frac{1}{\sqrt{2\pi(\sigma_L^2 + \sigma^2)}} \exp\left(-\frac{(\ell - L_0)^2}{2(\sigma_L^2 + \sigma^2)}\right).$$

Suppose we have two people, Alice and Bob, who agree on the values of σ and σ_L, but have different opinions about the value of L_0. Alice says it is L_{0A} (model M_1), Bob says it is L_{0B} (model M_2). They decide to settle their dispute by asking an (independent) student to read a new value ℓ. They would then work out the quantity

$$\ln \frac{\mathrm{prob}(M_1|\mathbf{y}, I)}{\mathrm{prob}(M_2|\mathbf{y}, I)} = \frac{(\ell - L_{0B})^2 - (\ell - L_{0A})^2}{2(\sigma_L^2 + \sigma^2)}$$

If the new value ℓ is closer to L_{0A}, then the fraction is positive making us prefer Alice's model M_1 over Bobs model M_2, and vice versa for ℓ being closer to L_{0B}. It is also interesting to see that the total variance $\sigma_L^2 + \sigma^2$ sets the scale on which the difference in the numerator is measured.

Second example. As a second example, consider the case in which Alice and Bob agree on L_0, but disagree on $\sigma_L^2 + \sigma^2$. Alice is more conservative and suggests a σ_A which is bigger than σ_B. In this case they would work out

$$\ln \frac{\mathrm{prob}(M_1|\mathbf{y}, I)}{\mathrm{prob}(M_2|\mathbf{y}, I)} = \ln \frac{\sigma_B}{\sigma_A} + \frac{(\ell - L_0)^2}{2} \left(\frac{1}{\sigma_B^2} - \frac{1}{\sigma_A^2}\right).$$

The second term on the right hand side is positive, and therefore always favors the more conservative model of Alice. The first term, however, gives Bob's model a little advantage. As a result, if

$$(\ell - L_0)^2 > \left(\frac{1}{\sigma_B^2} - \frac{1}{\sigma_A^2}\right)^{-1} \ln \frac{\sigma_A^2}{\sigma_B^2},$$

we would prefer Alices model over Bobs. Only if there is sufficient reason (in the sense that the new observation deviates significantly from L_0) do we have a reason to prefer the model with the larger uncertainty. Putting in numbers: suppose $\sigma_A = e\sigma_B$ ($e \approx 2.7$). Then we prefer Alices model, if $|\ell - L_0| > 1.5\sigma_B$. If $\sigma_A = 90\sigma_B$, then we would prefer Alice's model, if $|\ell - L_0| > 3\sigma_B$. These examples show that our tool for model comparison gives clear quantitative answers and leads to most reasonable results.

8.2 Beyond the evidence

Equation (53) can be further transformed into

$$\begin{aligned}\ln\frac{\mathrm{prob}(M_1|\mathbf{y},I)}{\mathrm{prob}(M_2|\mathbf{y},I)} &= \ln\frac{\mathrm{prob}(M_1|I)}{\mathrm{prob}(M_2|I)} + \ln\frac{\int d\boldsymbol{\theta}_1 \mathrm{prob}(\boldsymbol{\theta}_1,\mathbf{y}|M_1,I)}{\int d\boldsymbol{\theta}_2 \mathrm{prob}(\boldsymbol{\theta}_2,\mathbf{y}|M_2,I)} \\ &= \ln\frac{\mathrm{prob}(M_1|I)}{\mathrm{prob}(M_2|I)} + \ln\frac{\int d\boldsymbol{\theta}_1 \mathrm{prob}(\boldsymbol{\theta}_1|M_1,I)\mathrm{prob}(\mathbf{y}|\boldsymbol{\theta}_1,M_1,I)}{\int d\boldsymbol{\theta}_2 \mathrm{prob}(\boldsymbol{\theta}_2|M_2,I)\mathrm{prob}(\mathbf{y}|\boldsymbol{\theta}_2,M_2,I)}.\end{aligned}$$

We see that the probability products under the integrals are known from parameter estimation problems. Suppose now that we have determined $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ from the maximum likelihood estimate

$$\hat{\boldsymbol{\theta}}_i = \arg\min_{\boldsymbol{\theta}_i}\{-\ln\mathrm{prob}(\mathbf{y}|\boldsymbol{\theta}_i,M_i,I)\} \tag{55}$$

Then we may write the integrals as

$$\int d\boldsymbol{\theta}_i \mathrm{prob}(\boldsymbol{\theta}_i|M_i,I)\mathrm{prob}(\mathbf{y}|\boldsymbol{\theta}_i,M_i,I) = \\ \mathrm{prob}(\mathbf{y}|\hat{\boldsymbol{\theta}}_i,M_i,I)\int d\boldsymbol{\theta}_i \mathrm{prob}(\boldsymbol{\theta}_i|M_i,I)\frac{\mathrm{prob}(\mathbf{y}|\boldsymbol{\theta}_i,M_i,I)}{\mathrm{prob}(\mathbf{y}|\hat{\boldsymbol{\theta}}_i,M_i,I)}$$

giving

$$\begin{aligned}\ln\frac{\mathrm{prob}(M_1|\mathbf{y},I)}{\mathrm{prob}(M_2|\mathbf{y},I)} &= \ln\frac{\mathrm{prob}(M_1|I)}{\mathrm{prob}(M_2|I)} + \ln\frac{\mathrm{prob}(\mathbf{y}|\hat{\boldsymbol{\theta}}_1,M_1,I)}{\mathrm{prob}(\mathbf{y}|\hat{\boldsymbol{\theta}}_2,M_2,I)} \\ &+ \ln\frac{\int d\boldsymbol{\theta}_1 \mathrm{prob}(\boldsymbol{\theta}_1|M_1,I)\mathrm{prob}(\mathbf{y}|\boldsymbol{\theta}_1,M_1,I)/\mathrm{prob}(\mathbf{y}|\hat{\boldsymbol{\theta}}_1,M_1,I)}{\int d\boldsymbol{\theta}_2 \mathrm{prob}(\boldsymbol{\theta}_2|M_2,I)\mathrm{prob}(\mathbf{y}|\boldsymbol{\theta}_2,M_2,I)/\mathrm{prob}(\mathbf{y}|\hat{\boldsymbol{\theta}}_2,M_2,I)}.\end{aligned} \tag{56}$$

In this expression, the second term generates preference for the model with the larger maximum likelihood. The interpretation of the last term is less obvious. Its value does not only depend on the likelihood, but also on the width of the prior pdf, and on the number of parameters of each model.

8.3 Model comparison: an example

We will illustrate the above general concept by picking up an example from one of the previous lectures, where for the determination of the gravitational acceleration g, a set of data $\mathbf{y} = \{(z_i, t_i)\}$ with $N = 31$ points was measured. Two alternative fitting functions were used, namely

$$t_i = f_1(z_i; g) = \sqrt{\frac{2z_i}{g}} \quad \text{(model 1)},$$

with fit parameter g, and

$$t_i = f_2(z_i; g, z_0) = \sqrt{\frac{2(z_i - z_0)}{g}} \quad \text{(model 2)},$$

with fit parameters g and z_0. In both cases we used a least mean square fit implying the likelihood function for model 1

$$\mathrm{prob}(\mathbf{y}|g, \sigma, M, I) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{NQ_1(g)}{2\sigma^2}\right). \tag{57}$$

In case of model 2, $Q_1(g)$ is replaced by $Q_2(g, z_0)$.

If we give no prior preference to any of the two models, eq. (56) becomes after some calculations

$$\ln \frac{\mathrm{prob}(M_1|\mathbf{y}, I)}{\mathrm{prob}(M_2|\mathbf{y}, I)} = \frac{N-1}{2} \ln \frac{Q_{2,\mathrm{min}}}{Q_{1,\mathrm{min}}} + \ln\left(\frac{\sigma_z}{\sqrt{2Q_{2,\mathrm{min}}}} \sqrt{\frac{AB - C^2}{q}}\right) + \ln \frac{\Gamma((N-1)/2)}{\Gamma((N-2)/2)}.$$

The quantities A, B, and C are the elements of the Hesse matrix (the matrix of second derivatives) of $Q_2(g, z_0)$ evaluated at the minimum of this function. Correspondingly, the quantity q is the second derivative of $Q_1(g)$ evaluated at its minimum. The first term in this result favors the model that fits the data better. The second and third term, however, punish the use of the additional parameter z_0 in model 2. Numerically, the last term favors M_1 weakly with a value of 3.78. The first term, however, favors M_2 strongly with -73.89. We further find $\sqrt{AB - C^2} = 0.00905\,\mathrm{s}^4/\mathrm{m}^2$, $\sqrt{Q_{2,\mathrm{min}}} = 0.215\,\mathrm{ms}$, $\sqrt{q} = 0.02616\,\mathrm{s}^3/\mathrm{m}$. We see that the second term depends on the width σ_z that we assign to the prior pdf of z_0. Given our background information about the experiment, we can be quite sure that $\sigma_z = 1\,\mathrm{m}$ is certainly not too optimistic. Larger values can safely be rejected for an apparatus that is about 1 m in size. The second term therefore favors M_1 with a value of about +7.0, which is another punishment for the additional parameter. In total we have

$$\ln \frac{\text{prob}(M_1|\mathbf{y}, I)}{\text{prob}(M_2|\mathbf{y}, I)} = -73.89 + 7.0 + 3.78 = -63.11,$$

which implies that

$$\text{prob}(M_1|\mathbf{y}, I) = 3.9 \times 10^{-28} \text{prob}(M_2|\mathbf{y}, I).$$

This is a clear statement strongly favoring model 2 over model 1. We also see that this strong preference for M_2 would not be changed significantly by (unrealistic) higher values of σ_z, because the dependence is only logarithmic. However, it is also clear from the result that a smaller number of data points would have delivered much weaker evidence for M_2 as compared to M_1, because the decisive second term is directly proportional to N.

The fact that the second and third terms in the above result punish the use of a second parameter in model 2 may be seen as a mathematical implementation of 'Occam's razor'. This principle states, that if there are two competing hypotheses (models), the simpler should be preferred over the more complicated one, unless there is compelling evidence for the latter.