7.6 Bayes Factor

A very common example of the issues around model complexity is given by the simple polynomial fit : adding another term to a polynomial often gives a better fit (smaller chi - square per degree of freedom) but introduces an additional parameter.Use some random data (say, 20 numbers drawn from a Gaussian distribution) and the Bayes factor method to examine the odds in favour of adding one more polynomial term. Test that the Laplace approximation can be used to avoid a multidimensional numerical integration

Solution

Let's make this exercise fairly simple by assuming we are going to do a least-squares fit of the polynomials, assuming the residuals are Gaussian. Because the polynomials are linear in the coefficients we are looking for, it follows that the Laplace approximation will be close to exact. The outline of the reasoning is as follows.

Define the polynomial we are going to fit as

$$p(x, \vec{a}) = \sum_{i=1}^{N} a_i x^{i-1} \tag{1}$$

The Laplace approximation results from a Taylor series expansion of the logarithm of the likelihood function to second order. A single term in the likelihood, at the k-th data value y_k, is proportional to

$$\exp\left[-\frac{(y_k - p(x_k, \vec{a})^2}{2\sigma^2}\right] \tag{2}$$

as usual, in which the standard deviation in the data is σ. The total log likelihood then a summation over the M data values

$$\ln \mathcal{L} = \frac{1}{2\sigma^2} \sum_{k=1}^{M} (y_k - p(x_k, \vec{a})^2 \tag{3}$$

to within an additive constant that doesn't matter here.

Making a Taylor series expansion around the maximum likelihood (taken as a function of the parameters $\vec{a}$) we find that any higher order derivatives than

$$\frac{\partial^2 \ln \mathcal{L}}{\partial a_i \partial a_k}$$

will be zero because of the linearity of the function p in the as. So, the expansion to second order in the parameters in exact, and it follows that the likelihood function is exactly Gaussian in the as. This is the Laplace "approximation".

This is not quite enough in a Bayesian context, where we would like to apply the Laplace approximation to the posterior probability distribution: this allows us to integrate easily to get the Bayes factors. The posterior is of the form

$$\text{likelihood}(\vec{a}) \times \text{priors}(\vec{a})$$

and, as we have just seen, the first term is exactly Gaussian for the assumptions we have made. The width of this Gaussian will of course depend on the noise level in the data. If the priors are flat and significantly wider than the Gaussian, then the evidence calculated from the Laplace approximation should be pretty good.

Suppose for instance that the priors on the parameters are uniform, over some width for each parameter a_k of w_k. If we are in a situation where we expect the data to affect our prior beliefs materially, then obviously we must have each w_k appreciably larger than the width of the likelihood function, or else our prior opinions could not be affected much by the data. In this case, the cut-offs in the uniform distributions assumed for the priors cannot affect the integration over the likelihood function very much and so Bayes factors, calculated from the Laplace approximation, must be good approximations.

The Laplace approximation need not be a good one; Figure 6.5 in the text shows an example where the approximation is not good in detail. In this case the model is not linear in the parameters although a Gaussian distribution of residuals was still assumed.

On the basis of this analysis, we will assume that the Laplace approximation is sufficient for the purposes of this example. In the next example we will do similar calculations but use numerical MCMC methods.

The solution then involves the following steps.

1. Generation of some simulated data. Two realizations are given in the data files for this example. The standard deviation on each data point is $\sigma = 0.1$.

2. Choice of fitting function. A cubic and a quartic polynomial were used, as in Equation 1.

3. Statistics and priors. The residuals are assumed to be Gaussian, so that the likelihood function is given by Equation 3. The priors are

assumed to be sufficiently diffuse for the likelihood to be a good approximation to the posterior probability distribution; we will return to this point later.

4. The maximum of the likelihood is found numerically by a Newton-Raphson method, and the maximum likelihood estimates of the polynomial coefficients are

   (cubic) 0.188683, -0.0124804, 0.00402867, -0.000204107

   (quartic) 0.363918, -0.149563, 0.0317681, -0.00222427, 0.0000480991

   (A matrix inversion method would also work: see *Press et al., Numerical Recipes*.)

   From these we can calculate the value of the maximum likelihood $\mathcal{L}_{\max}$. Equation 3 gives the formula to within an additive constant, which is the same for both our polynomial models and so can be ignored. (It is the product of the normalizing factors of each of the terms of the form shown in Equation 2.)

5. The Bayes Factor. For either polynomial, the Bayes factor is of the form

$$\int_{-\infty}^{\infty} \mathcal{L}(\vec{a})\mathrm{pr}(\vec{a})\, d\vec{a}$$

   and, as discussed, we will ignore the priors pr in this integral (although the numerical normalizing factors for the priors will need to be accounted for). The Laplace approximation to the Bayes factor is then

$$\frac{(2\pi)^{N/2}}{\sqrt{|\mathcal{H}|}}\, \mathcal{L}_{\max}. \tag{4}$$

   $\mathcal{H}$, the Hessian matrix or matrix of second derivatives of $\ln \mathcal{L}$, contains some formidable numbers: this is why high-order polynomial fitting by least squares requires some care – again, see *Numerical Recipes*. This example is for the cubic case. Note it depends only on the values of the independent variable (x) that occur in the data, not on the dependent variable or the parameters – this is because we have a linear least squares model.

$$\mathcal{H} = \frac{1}{\sigma^2}\begin{pmatrix} 20 & 210 & 2870 & 44100 & 722666 \\ 210 & 2870 & 44100 & 722666 & 12333300 \\ 2870 & 44100 & 722666 & 12333300 & 216455810 \\ 44100 & 722666 & 12333300 & 216455810 & 3877286700 \\ 722666 & 12333300 & 216455810 & 3877286700 & 70540730666 \end{pmatrix}$$

The ratio of the terms given in 4, computed for the cubic and quartic case, is the Bayes factor, except for the priors. If the priors on the parameters are the same in the cubic and quartic cases, then all of their normalizing factors cancel out except for the normalizing factor associated with the quartic term, which does not appear in the cubic model at all. We have assumed that the priors on the polynomial coefficients a_k are uniformly distributed with widths w_k, so the normalizing factor for the prior on the quartic coefficient a_5 is $1/w_5$. The Bayes factor is therefore

$$\mathcal{B} = w_5 \frac{\frac{(2\pi)^{4/2}}{\sqrt{|\mathcal{H}_{cubic}|}}\,\mathcal{L}_{max,cubic}}{\frac{(2\pi)^{5/2}}{\sqrt{|\mathcal{H}_{quartic}|}}\,\mathcal{L}_{max,quartic}}.$$

6. Results. For the two datasets given, the values of the Bayes factor are

$$0.9 \times 10^4\, w_5$$

$$\tfrac{1}{12.6} \times 10^4\, w_5.$$

We see that the prior on the quartic term is crucial to the result. Now the priors are specific to the problem to hand, but for the present example we might make this argument. The quartic term is $a_5 x^4$. The range of the data is ± 10 and the dependent variable gets no bigger than 1, so *a priori* we have $a_5 10^4 < 1$, meaning that a_5 is uniformly distributed within $|a_5| < 10^{-4}$. This (rough) argument tells us that $w_5 \simeq 10^{-4}$. Assuming this value, we find then that the odds on the cubic fit for our first example (Figure 1) are 0.9 to 1, and for the second example (Figure 2), the odds on the quartic are 12.6 to 1. These are not large odds either way, which is what one would expect from rather structureless data, but they are in the right sense in each case.

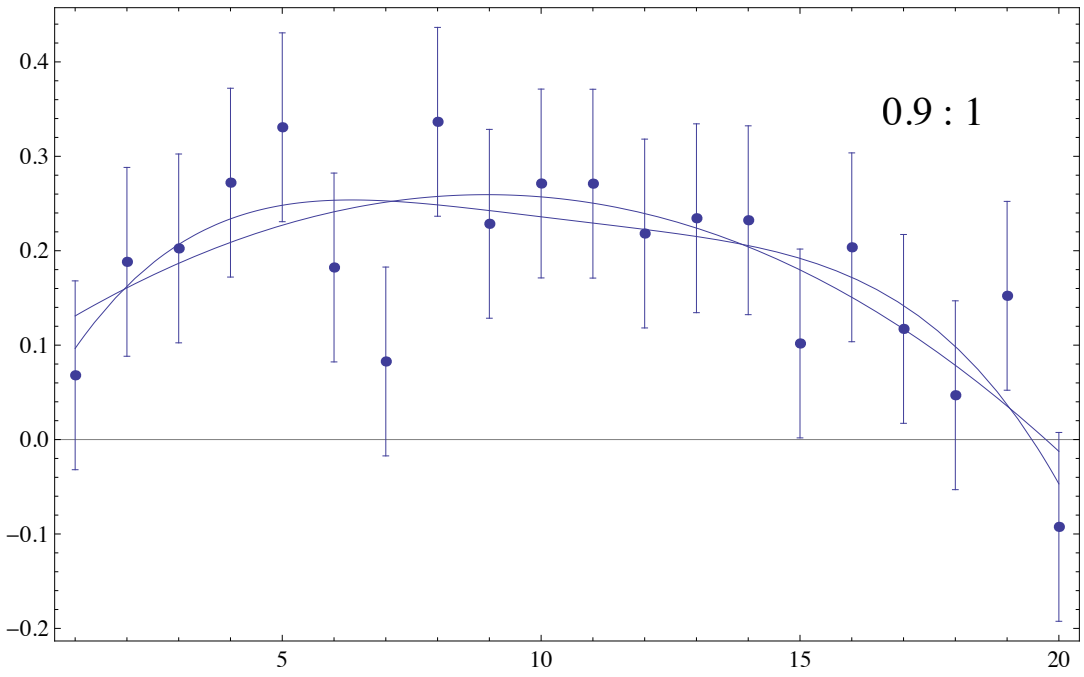

Figure 1: Simulated data with cubic and quartic fits shown. The fits are very similar and the odds are weakly in favour of the simpler model.

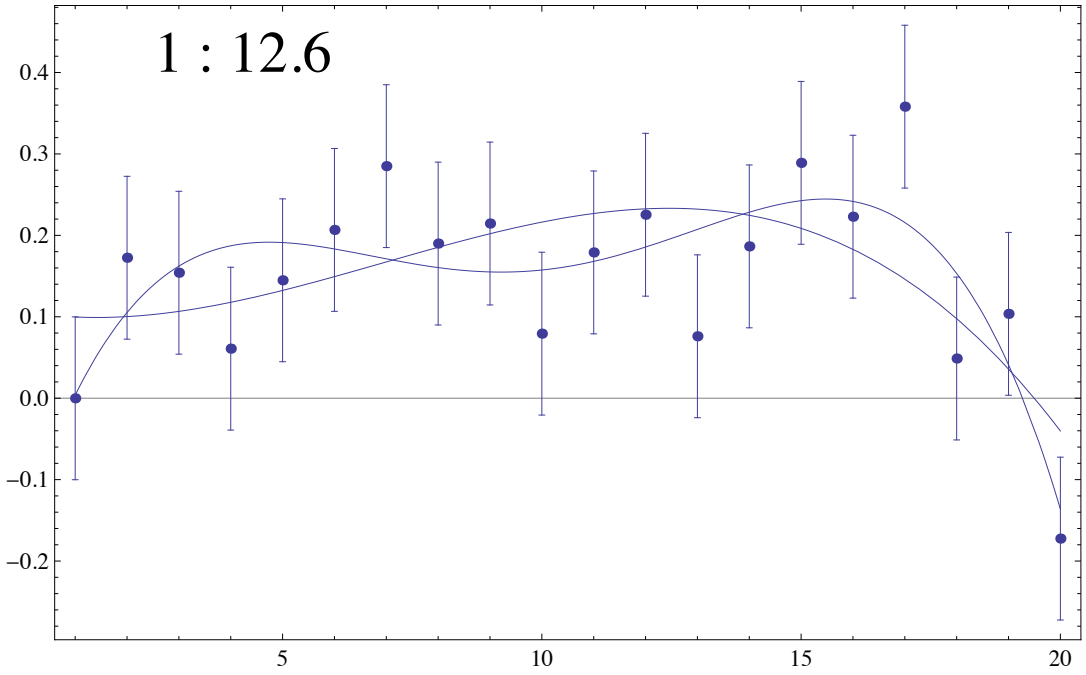

Figure 2: Simulated data with cubic and quartic fits shown. The fits are different, with the quartic taking up more of the variation in the data. Despite being a more complex model the odds are mildly in its favour.

The key role of the prior on the extra (quartic) parameter is very instructive. This prior largely determines the Ockham factor (the determinant of the Hessian matrix also plays a role).