

28

Model Comparison and Occam's Razor

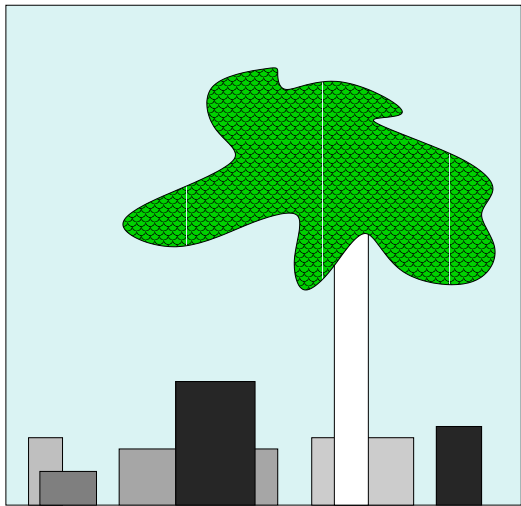


Figure 28.1. A picture to be interpreted. It contains a tree and some boxes.

► 28.1 Occam's razor

How many boxes are in the picture (figure 28.1)? In particular, how many boxes are in the vicinity of the tree? If we looked with x-ray spectacles, would we see one or two boxes behind the trunk (figure 28.2)? (Or even more?) Occam's razor is the principle that states a preference for simple theories. 'Accept the simplest explanation that fits the data'. Thus according to Occam's razor, we should deduce that there is only one box behind the tree. Is this an ad hoc rule of thumb? Or is there a convincing reason for believing there is most likely one box? Perhaps your intuition likes the argument 'well, it would be a remarkable *coincidence* for the two boxes to be just the same height and colour as each other'. If we wish to make artificial intelligences that interpret data correctly, we must translate this intuitive feeling into a concrete theory.

Motivations for Occam's razor

If several explanations are compatible with a set of observations, Occam's razor advises us to buy the simplest. This principle is often advocated for one of two reasons: the first is aesthetic ('A theory with mathematical beauty is more likely to be correct than an ugly one that fits some experimental data')

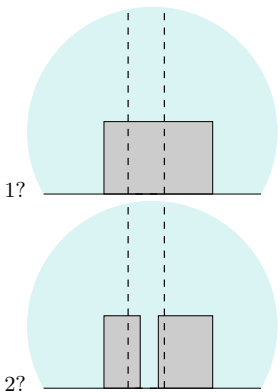
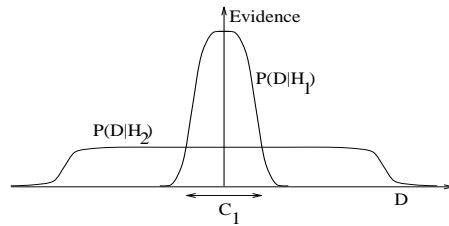


Figure 28.2. How many boxes are behind the tree?



(Paul Dirac)); the second reason is the past empirical success of Occam's razor. However there is a different justification for Occam's razor, namely:

Coherent inference (as embodied by Bayesian probability) automatically embodies Occam's razor, quantitatively.

It is indeed *more probable* that there's one box behind the tree, and we can compute how much more probable one is than two.

Model comparison and Occam's razor

We evaluate the plausibility of two alternative theories \mathcal{H}_1 and \mathcal{H}_2 in the light of data D as follows: using Bayes' theorem, we relate the plausibility of model \mathcal{H}_1 given the data, $P(\mathcal{H}_1 | D)$, to the predictions made by the model about the data, $P(D | \mathcal{H}_1)$, and the prior plausibility of \mathcal{H}_1 , $P(\mathcal{H}_1)$. This gives the following probability ratio between theory \mathcal{H}_1 and theory \mathcal{H}_2 :

$$\frac{P(\mathcal{H}_1 | D)}{P(\mathcal{H}_2 | D)} = \frac{P(\mathcal{H}_1) P(D | \mathcal{H}_1)}{P(\mathcal{H}_2) P(D | \mathcal{H}_2)}. \quad (28.1)$$

The first ratio ($P(\mathcal{H}_1)/P(\mathcal{H}_2)$) on the right-hand side measures how much our initial beliefs favoured \mathcal{H}_1 over \mathcal{H}_2 . The second ratio expresses how well the observed data were predicted by \mathcal{H}_1 , compared to \mathcal{H}_2 .

How does this relate to Occam's razor, when \mathcal{H}_1 is a simpler model than \mathcal{H}_2 ? The first ratio ($P(\mathcal{H}_1)/P(\mathcal{H}_2)$) gives us the opportunity, if we wish, to insert a prior bias in favour of \mathcal{H}_1 on aesthetic grounds, or on the basis of experience. This would correspond to the aesthetic and empirical motivations for Occam's razor mentioned earlier. But such a prior bias is not necessary: the second ratio, the data-dependent factor, embodies Occam's razor *automatically*. Simple models tend to make precise predictions. Complex models, by their nature, are capable of making a greater variety of predictions (figure 28.3). So if \mathcal{H}_2 is a more complex model, it must spread its predictive probability $P(D | \mathcal{H}_2)$ more thinly over the data space than \mathcal{H}_1 . Thus, in the case where the data are compatible with both theories, the simpler \mathcal{H}_1 will turn out more probable than \mathcal{H}_2 , without our having to express any subjective dislike for complex models. Our subjective prior just needs to assign equal prior probabilities to the possibilities of simplicity and complexity. Probability theory then allows the observed data to express their opinion.

Let us turn to a simple example. Here is a sequence of numbers:

$$-1, 3, 7, 11.$$

The task is to predict the next two numbers, and infer the underlying process that gave rise to this sequence. A popular answer to this question is the prediction '15, 19', with the explanation 'add 4 to the previous number'.

What about the alternative answer '−19.9, 1043.8' with the underlying rule being: 'get the next number from the previous number, x , by evaluating

Figure 28.3. Why Bayesian inference embodies Occam's razor. This figure gives the basic intuition for why complex models can turn out to be less probable. The horizontal axis represents the space of possible data sets D . Bayes' theorem rewards models in proportion to how much they *predicted* the data that occurred. These predictions are quantified by a normalized probability distribution on D . This probability of the data given model \mathcal{H}_i , $P(D | \mathcal{H}_i)$, is called the evidence for \mathcal{H}_i . A simple model \mathcal{H}_1 makes only a limited range of predictions, shown by $P(D | \mathcal{H}_1)$; a more powerful model \mathcal{H}_2 , that has, for example, more free parameters than \mathcal{H}_1 , is able to predict a greater variety of data sets. This means, however, that \mathcal{H}_2 does not predict the data sets in region C_1 as strongly as \mathcal{H}_1 . Suppose that equal prior probabilities have been assigned to the two models. Then, if the data set falls in region C_1 , the *less powerful* model \mathcal{H}_1 will be the *more probable* model.

$-x^3/11 + 9/11x^2 + 23/11$? I assume that this prediction seems rather less plausible. But the second rule fits the data $(-1, 3, 7, 11)$ just as well as the rule 'add 4'. So why should we find it less plausible? Let us give labels to the two general theories:

\mathcal{H}_a – the sequence is an *arithmetic* progression, 'add n ', where n is an integer.

\mathcal{H}_c – the sequence is generated by a *cubic* function of the form $x \rightarrow cx^3 + dx^2 + e$, where c, d and e are fractions.

One reason for finding the second explanation, \mathcal{H}_c , less plausible, might be that arithmetic progressions are more frequently encountered than cubic functions. This would put a bias in the prior probability ratio $P(\mathcal{H}_a)/P(\mathcal{H}_c)$ in equation (28.1). But let us give the two theories equal prior probabilities, and concentrate on what the data have to say. How well did each theory predict the data?

To obtain $P(D|\mathcal{H}_a)$ we must specify the probability distribution that each model assigns to its parameters. First, \mathcal{H}_a depends on the added integer n , and the first number in the sequence. Let us say that these numbers could each have been anywhere between -50 and 50 . Then since only the pair of values $\{n=4, \text{first number} = -1\}$ give rise to the observed data $D = (-1, 3, 7, 11)$, the probability of the data, given \mathcal{H}_a , is:

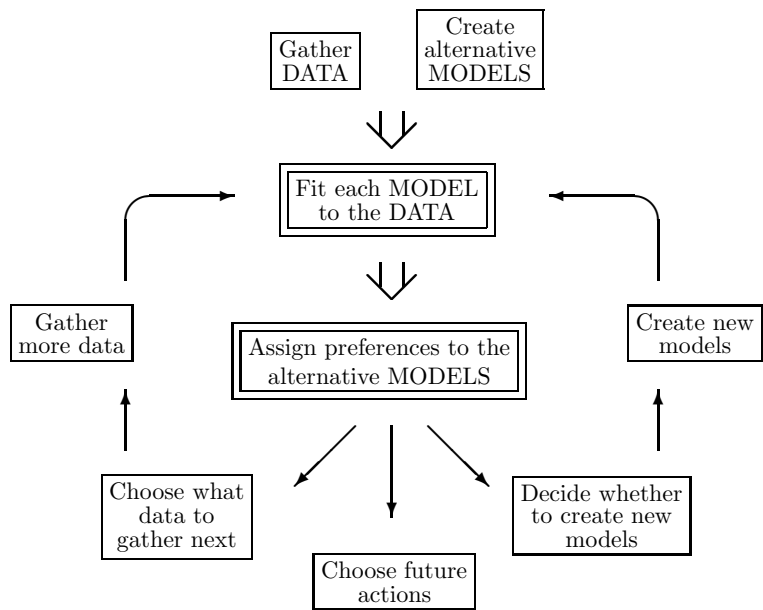
$$P(D|\mathcal{H}_a) = \frac{1}{101} \frac{1}{101} = 0.00010. \quad (28.2)$$

To evaluate $P(D|\mathcal{H}_c)$, we must similarly say what values the fractions c, d and e might take on. [I choose to represent these numbers as fractions rather than real numbers because if we used real numbers, the model would assign, relative to \mathcal{H}_a , an infinitesimal probability to D . Real parameters are the norm however, and are assumed in the rest of this chapter.] A reasonable prior might state that for each fraction the numerator could be any number between -50 and 50 , and the denominator is any number between 1 and 50 . As for the initial value in the sequence, let us leave its probability distribution the same as in \mathcal{H}_a . There are four ways of expressing the fraction $c = -1/11 = -2/22 = -3/33 = -4/44$ under this prior, and similarly there are four and two possible solutions for d and e , respectively. So the probability of the observed data, given \mathcal{H}_c , is found to be:

$$\begin{aligned} P(D|\mathcal{H}_c) &= \left(\frac{1}{101}\right) \left(\frac{4}{101} \frac{1}{50}\right) \left(\frac{4}{101} \frac{1}{50}\right) \left(\frac{2}{101} \frac{1}{50}\right) \\ &= 0.0000000000025 = 2.5 \times 10^{-12}. \end{aligned} \quad (28.3)$$

Thus comparing $P(D|\mathcal{H}_c)$ with $P(D|\mathcal{H}_a) = 0.00010$, even if our prior probabilities for \mathcal{H}_a and \mathcal{H}_c are equal, the odds, $P(D|\mathcal{H}_a) : P(D|\mathcal{H}_c)$, in favour of \mathcal{H}_a over \mathcal{H}_c , given the sequence $D = (-1, 3, 7, 11)$, are about forty million to one. \square

This answer depends on several subjective assumptions; in particular, the probability assigned to the free parameters n, c, d, e of the theories. Bayesians make no apologies for this: there is no such thing as inference or prediction without assumptions. However, the quantitative details of the prior probabilities have no effect on the qualitative Occam's razor effect; the complex theory \mathcal{H}_c always suffers an 'Occam factor' because it has more parameters, and so can predict a greater variety of data sets (figure 28.3). This was only a small example, and there were only four data points; as we move to larger



and more sophisticated problems the magnitude of the Occam factors typically increases, and the degree to which our inferences are influenced by the quantitative details of our subjective assumptions becomes smaller.

Bayesian methods and data analysis

Let us now relate the discussion above to real problems in data analysis.

There are countless problems in science, statistics and technology which require that, given a limited data set, preferences be assigned to alternative models of differing complexities. For example, two alternative hypotheses accounting for planetary motion are Mr. Inquisition's geocentric model based on 'epicycles', and Mr. Copernicus's simpler model of the solar system with the sun at the centre. The epicyclic model fits data on planetary motion at least as well as the Copernican model, but does so using more parameters. Coincidentally for Mr. Inquisition, two of the extra epicyclic parameters for every planet are found to be identical to the period and radius of the sun's 'cycle around the earth'. Intuitively we find Mr. Copernicus's theory more probable.

The mechanism of the Bayesian razor: the evidence and the Occam factor

Two levels of inference can often be distinguished in the process of data modelling. At the first level of inference, we assume that a particular model is true, and we fit that model to the data, i.e., we infer what values its free parameters should plausibly take, given the data. The results of this inference are often summarized by the most probable parameter values, and error bars on those parameters. This analysis is repeated for each model. The second level of inference is the task of model comparison. Here we wish to compare the models in the light of the data, and assign some sort of preference or ranking to the alternatives.

Note that both levels of *inference* are distinct from *decision theory*. The goal of inference is, given a defined hypothesis space and a particular data set, to assign probabilities to hypotheses. Decision theory typically chooses between alternative *actions* on the basis of these probabilities so as to minimize the

Figure 28.4. Where Bayesian inference fits into the data modelling process. This figure illustrates an abstraction of the part of the scientific process in which data are collected and modelled. In particular, this figure applies to pattern classification, learning, interpolation, etc. The two double-framed boxes denote the two steps which involve *inference*. It is only in those two steps that Bayes' theorem can be used. Bayes does not tell you how to invent models, for example. The first box, 'fitting each model to the data', is the task of inferring what the model parameters might be given the model and the data. Bayesian methods may be used to find the most probable parameter values, and error bars on those parameters. The result of applying Bayesian methods to this problem is often little different from the answers given by orthodox statistics. The second inference task, model comparison in the light of the data, is where Bayesian methods are in a class of their own. This second inference problem requires a quantitative Occam's razor to penalize over-complex models. Bayesian methods can assign objective preferences to the alternative models in a way that automatically embodies Occam's razor.

expectation of a 'loss function'. This chapter concerns inference alone and no loss functions are involved. When we discuss model comparison, this should not be construed as implying model *choice*. Ideal Bayesian predictions do not involve choice between models; rather, predictions are made by summing over all the alternative models, weighted by their probabilities.

Bayesian methods are able consistently and quantitatively to solve both the inference tasks. There is a popular myth that states that Bayesian methods differ from orthodox statistical methods only by the inclusion of subjective priors, which are difficult to assign, and which usually don't make much difference to the conclusions. It is true that, at the first level of inference, a Bayesian's results will often differ little from the outcome of an orthodox attack. What is not widely appreciated is how a Bayesian performs the second level of inference; this chapter will therefore focus on Bayesian model comparison.

Model comparison is a difficult task because it is not possible simply to choose the model that fits the data best: more complex models can always fit the data better, so the maximum likelihood model choice would lead us inevitably to implausible, over-parameterized models, which generalize poorly. Occam's razor is needed.

Let us write down Bayes' theorem for the two levels of inference described above, so as to see explicitly how Bayesian model comparison works. Each model \mathcal{H}_i is assumed to have a vector of parameters \mathbf{w} . A model is defined by a collection of probability distributions: a 'prior' distribution $P(\mathbf{w} | \mathcal{H}_i)$, which states what values the model's parameters might be expected to take; and a set of conditional distributions, one for each value of \mathbf{w} , defining the predictions $P(D | \mathbf{w}, \mathcal{H}_i)$ that the model makes about the data D .

1. **Model fitting.** At the first level of inference, we assume that one model, the i th, say, is true, and we infer what the model's parameters \mathbf{w} might be, given the data D . Using Bayes' theorem, the *posterior probability* of the parameters \mathbf{w} is:

$$P(\mathbf{w} | D, \mathcal{H}_i) = \frac{P(D | \mathbf{w}, \mathcal{H}_i)P(\mathbf{w} | \mathcal{H}_i)}{P(D | \mathcal{H}_i)}, \quad (28.4)$$

that is,

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}.$$

The normalizing constant $P(D | \mathcal{H}_i)$ is commonly ignored since it is irrelevant to the first level of inference, i.e., the inference of \mathbf{w} ; but it becomes important in the second level of inference, and we name it the *evidence* for \mathcal{H}_i . It is common practice to use gradient-based methods to find the maximum of the posterior, which defines the most probable value for the parameters, \mathbf{w}_{MP} ; it is then usual to summarize the posterior distribution by the value of \mathbf{w}_{MP} , and error bars or confidence intervals on these best-fit parameters. Error bars can be obtained from the curvature of the posterior; evaluating the Hessian at \mathbf{w}_{MP} , $\mathbf{A} = -\nabla \nabla \ln P(\mathbf{w} | D, \mathcal{H}_i)|_{\mathbf{w}_{\text{MP}}}$, and Taylor-expanding the log posterior probability with $\Delta \mathbf{w} = \mathbf{w} - \mathbf{w}_{\text{MP}}$:

$$P(\mathbf{w} | D, \mathcal{H}_i) \simeq P(\mathbf{w}_{\text{MP}} | D, \mathcal{H}_i) \exp \left(-1/2 \Delta \mathbf{w}^T \mathbf{A} \Delta \mathbf{w} \right), \quad (28.5)$$

we see that the posterior can be locally approximated as a Gaussian with covariance matrix (equivalent to error bars) \mathbf{A}^{-1} . [Whether this approximation is good or not will depend on the problem we are solving. Indeed, the maximum and mean of the posterior distribution have

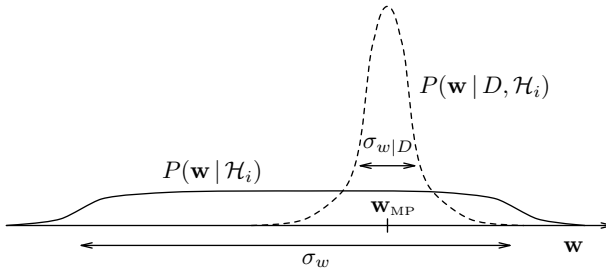


Figure 28.5. The Occam factor. This figure shows the quantities that determine the Occam factor for a hypothesis \mathcal{H}_i having a single parameter \mathbf{w} . The prior distribution (solid line) for the parameter has width σ_w . The posterior distribution (dashed line) has a single peak at \mathbf{w}_{MP} with characteristic width $\sigma_{w|D}$. The Occam factor is

$$\sigma_{w|D} P(\mathbf{w}_{\text{MP}} | \mathcal{H}_i) = \frac{\sigma_w | D}{\sigma_w}.$$

no fundamental status in Bayesian inference – they both change under nonlinear reparameterizations. Maximization of a posterior probability is useful only if an approximation like equation (28.5) gives a good summary of the distribution.]

2. **Model comparison.** At the second level of inference, we wish to infer which model is most plausible given the data. The posterior probability of each model is:

$$P(\mathcal{H}_i | D) \propto P(D | \mathcal{H}_i) P(\mathcal{H}_i). \quad (28.6)$$

Notice that the data-dependent term $P(D | \mathcal{H}_i)$ is the evidence for \mathcal{H}_i , which appeared as the normalizing constant in (28.4). The second term, $P(\mathcal{H}_i)$, is the subjective prior over our hypothesis space, which expresses how plausible we thought the alternative models were before the data arrived. Assuming that we choose to assign equal priors $P(\mathcal{H}_i)$ to the alternative models, *models \mathcal{H}_i are ranked by evaluating the evidence*. The normalizing constant $P(D) = \sum_i P(D | \mathcal{H}_i) P(\mathcal{H}_i)$ has been omitted from equation (28.6) because in the data-modelling process we may develop new models after the data have arrived, when an inadequacy of the first models is detected, for example. Inference is open ended: we continually seek more probable models to account for the data we gather.

To repeat the key idea: to rank alternative models \mathcal{H}_i , a Bayesian evaluates the evidence $P(D | \mathcal{H}_i)$. This concept is very general: the evidence can be evaluated for parametric and ‘non-parametric’ models alike; whatever our data-modelling task, a regression problem, a classification problem, or a density estimation problem, the evidence is a transportable quantity for comparing alternative models. In all these cases the evidence naturally embodies Occam’s razor.

Evaluating the evidence

Let us now study the evidence more closely to gain insight into how the Bayesian Occam’s razor works. The evidence is the normalizing constant for equation (28.4):

$$P(D | \mathcal{H}_i) = \int P(D | \mathbf{w}, \mathcal{H}_i) P(\mathbf{w} | \mathcal{H}_i) d\mathbf{w}. \quad (28.7)$$

For many problems the posterior $P(\mathbf{w} | D, \mathcal{H}_i) \propto P(D | \mathbf{w}, \mathcal{H}_i) P(\mathbf{w} | \mathcal{H}_i)$ has a strong peak at the most probable parameters \mathbf{w}_{MP} (figure 28.5). Then, taking for simplicity the one-dimensional case, the evidence can be approximated, using Laplace’s method, by the height of the peak of the integrand $P(D | \mathbf{w}, \mathcal{H}_i) P(\mathbf{w} | \mathcal{H}_i)$ times its width, $\sigma_{w|D}$:

$$P(D | \mathcal{H}_i) \simeq \underbrace{P(D | \mathbf{w}_{\text{MP}}, \mathcal{H}_i)}_{\text{Best fit likelihood}} \times \underbrace{P(\mathbf{w}_{\text{MP}} | \mathcal{H}_i) \sigma_{w|D}}_{\text{Occam factor}}. \quad (28.8)$$

Evidence \simeq Best fit likelihood \times Occam factor

Thus the evidence is found by taking the best-fit likelihood that the model can achieve and multiplying it by an 'Occam factor', which is a term with magnitude less than one that penalizes \mathcal{H}_i for having the parameter \mathbf{w} .

Interpretation of the Occam factor

The quantity $\sigma_{w|D}$ is the posterior uncertainty in \mathbf{w} . Suppose for simplicity that the prior $P(\mathbf{w} | \mathcal{H}_i)$ is uniform on some large interval σ_w , representing the range of values of \mathbf{w} that were possible *a priori*, according to \mathcal{H}_i (figure 28.5). Then $P(\mathbf{w}_{\text{MP}} | \mathcal{H}_i) = 1/\sigma_w$, and

$$\text{Occam factor} = \frac{\sigma_{w|D}}{\sigma_w}, \quad (28.9)$$

i.e., the Occam factor is equal to the ratio of the posterior accessible volume of \mathcal{H}_i 's parameter space to the prior accessible volume, or the factor by which \mathcal{H}_i 's hypothesis space collapses when the data arrive. The model \mathcal{H}_i can be viewed as consisting of a certain number of exclusive submodels, of which only one survives when the data arrive. The Occam factor is the inverse of that number. The logarithm of the Occam factor is a measure of the amount of information we gain about the model's parameters when the data arrive.

A complex model having many parameters, each of which is free to vary over a large range σ_w , will typically be penalized by a stronger Occam factor than a simpler model. The Occam factor also penalizes models that have to be finely tuned to fit the data, favouring models for which the required precision of the parameters $\sigma_{w|D}$ is coarse. The magnitude of the Occam factor is thus a measure of complexity of the model; it relates to the complexity of the predictions that the model makes in data space. This depends not only on the number of parameters in the model, but also on the prior probability that the model assigns to them. Which model achieves the greatest evidence is determined by a trade-off between minimizing this natural complexity measure and minimizing the data misfit. In contrast to alternative measures of model complexity, the Occam factor for a model is straightforward to evaluate: it simply depends on the error bars on the parameters, which we already evaluated when fitting the model to the data.

Figure 28.6 displays an entire hypothesis space so as to illustrate the various probabilities in the analysis. There are three models, $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3$, which have equal prior probabilities. Each model has one parameter \mathbf{w} (each shown on a horizontal axis), but assigns a different prior range σ_w to that parameter. \mathcal{H}_3 is the most 'flexible' or 'complex' model, assigning the broadest prior range. A one-dimensional data space is shown by the vertical axis. Each model assigns a joint probability distribution $P(D, \mathbf{w} | \mathcal{H}_i)$ to the data and the parameters, illustrated by a cloud of dots. These dots represent random samples from the full probability distribution. The total number of dots in each of the three model subspaces is the same, because we assigned equal prior probabilities to the models.

When a particular data set D is received (horizontal line), we infer the posterior distribution of \mathbf{w} for a model (\mathcal{H}_3 , say) by reading out the density along that horizontal line, and normalizing. The posterior probability $P(\mathbf{w} | D, \mathcal{H}_3)$ is shown by the dotted curve at the bottom. Also shown is the prior distribution $P(\mathbf{w} | \mathcal{H}_3)$ (cf. figure 28.5). [In the case of model \mathcal{H}_1 which is very poorly

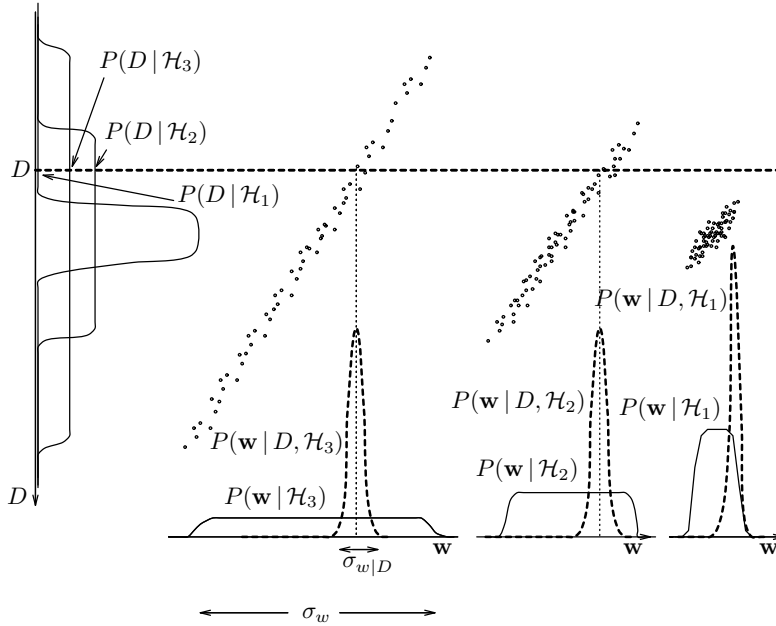


Figure 28.6. A hypothesis space consisting of three exclusive models, each having one parameter \mathbf{w} , and a one-dimensional data set D . The ‘data set’ is a single measured value which differs from the parameter \mathbf{w} by a small amount of additive noise. Typical samples from the joint distribution $P(D, \mathbf{w}, \mathcal{H})$ are shown by dots. (N.B., these are not data points.) The observed ‘data set’ is a single particular value for D shown by the dashed horizontal line. The dashed curves below show the posterior probability of \mathbf{w} for each model given this data set (cf. figure 28.3). The evidence for the different models is obtained by marginalizing onto the D axis at the left-hand side (cf. figure 28.5).

matched to the data, the shape of the posterior distribution will depend on the details of the tails of the prior $P(\mathbf{w} | \mathcal{H}_1)$ and the likelihood $P(D | \mathbf{w}, \mathcal{H}_1)$; the curve shown is for the case where the prior falls off more strongly.]

We obtain figure 28.3 by marginalizing the joint distributions $P(D, \mathbf{w} | \mathcal{H}_i)$ onto the D axis at the left-hand side. For the data set D shown by the dotted horizontal line, the evidence $P(D | \mathcal{H}_3)$ for the more flexible model \mathcal{H}_3 has a smaller value than the evidence for \mathcal{H}_2 . This is because \mathcal{H}_3 placed less predictive probability (fewer dots) on that line. In terms of the distributions over \mathbf{w} , model \mathcal{H}_3 has smaller evidence because the Occam factor $\sigma_{w|D}/\sigma_w$ is smaller for \mathcal{H}_3 than for \mathcal{H}_2 . The simplest model \mathcal{H}_1 has the smallest evidence of all, because the best fit that it can achieve to the data D is very poor. Given this data set, the most probable model is \mathcal{H}_2 .

Occam factor for several parameters

If the posterior is well approximated by a Gaussian, then the Occam factor is obtained from the determinant of the corresponding covariance matrix (cf. equation (28.8) and Chapter 27):

$$P(D | \mathcal{H}_i) \simeq \underbrace{P(D | \mathbf{w}_{\text{MP}}, \mathcal{H}_i)}_{\text{Evidence}} \times \underbrace{P(\mathbf{w}_{\text{MP}} | \mathcal{H}_i) \det^{-\frac{1}{2}}(\mathbf{A}/2\pi)}_{\text{Occam factor}}, \quad (28.10)$$

where $\mathbf{A} = -\nabla\nabla \ln P(\mathbf{w} | D, \mathcal{H}_i)$, the Hessian which we evaluated when we calculated the error bars on \mathbf{w}_{MP} (equation 28.5 and Chapter 27). As the amount of data collected increases, this Gaussian approximation is expected to become increasingly accurate.

In summary, Bayesian model comparison is a simple extension of maximum likelihood model selection: *the evidence is obtained by multiplying the best-fit likelihood by the Occam factor.*

To evaluate the Occam factor we need only the Hessian \mathbf{A} , if the Gaussian approximation is good. Thus the Bayesian method of model comparison by evaluating the evidence is no more computationally demanding than the task of finding for each model the best-fit parameters and their error bars.

► 28.2 Example

Let's return to the example that opened this chapter. Are there one or two boxes behind the tree in figure 28.1? Why do coincidences make us suspicious?

Let's assume the image of the area round the trunk and box has a size of 50 pixels, that the trunk is 10 pixels wide, and that 16 different colours of boxes can be distinguished. The theory \mathcal{H}_1 that says there is one box near the trunk has four free parameters: three coordinates defining the top three edges of the box, and one parameter giving the box's colour. (If boxes could levitate, there would be five free parameters.)

The theory \mathcal{H}_2 that says there are two boxes near the trunk has eight free parameters (twice four), plus a ninth, a binary variable that indicates which of the two boxes is the closest to the viewer.

What is the evidence for each model? We'll do \mathcal{H}_1 first. We need a prior on the parameters to evaluate the evidence. For convenience, let's work in pixels. Let's assign a separable prior to the horizontal location of the box, its width, its height, and its colour. The height could have any of, say, 20 distinguishable values, so could the width, and so could the location. The colour could have any of 16 values. We'll put uniform priors over these variables. We'll ignore all the parameters associated with other objects in the image, since they don't come into the model comparison between \mathcal{H}_1 and \mathcal{H}_2 . The evidence is

$$P(D|\mathcal{H}_1) = \frac{1}{20} \frac{1}{20} \frac{1}{20} \frac{1}{16} \quad (28.11)$$

since only one setting of the parameters fits the data, and it predicts the data perfectly.

As for model \mathcal{H}_2 , six of its nine parameters are well-determined, and three of them are partly-constrained by the data. If the left-hand box is furthest away, for example, then its width is at least 8 pixels and at most 30; if it's the closer of the two boxes, then its width is between 8 and 18 pixels. (I'm assuming here that the visible portion of the left-hand box is about 8 pixels wide.) To get the evidence we need to sum up the prior probabilities of all viable hypotheses. To do an exact calculation, we need to be more specific about the data and the priors, but let's just get the ballpark answer, assuming that the two unconstrained real variables have half their values available, and that the binary variable is completely undetermined. (As an exercise, you can make an explicit model and work out the exact answer.)

$$P(D|\mathcal{H}_2) \simeq \frac{1}{20} \frac{1}{20} \frac{10}{20} \frac{1}{16} \frac{1}{20} \frac{1}{20} \frac{10}{20} \frac{1}{16} \frac{2}{2}. \quad (28.12)$$

Thus the posterior probability ratio is (assuming equal prior probability):

$$\frac{P(D|\mathcal{H}_1)P(\mathcal{H}_1)}{P(D|\mathcal{H}_2)P(\mathcal{H}_2)} = \frac{1}{\frac{1}{20} \frac{10}{20} \frac{10}{20} \frac{1}{16}} \quad (28.13)$$

$$= 20 \times 2 \times 2 \times 16 \simeq 1000/1. \quad (28.14)$$

So the data are roughly 1000 to 1 in favour of the simpler hypothesis. The four factors in (28.13) can be interpreted in terms of Occam factors. The more complex model has four extra parameters for sizes and colours – three for sizes, and one for colour. It has to pay two big Occam factors ($1/20$ and $1/16$) for the highly suspicious coincidences that the two box heights match exactly and the two colours match exactly; and it also pays two lesser Occam factors for the two lesser coincidences that both boxes happened to have one of their edges conveniently hidden behind a tree or behind each other.

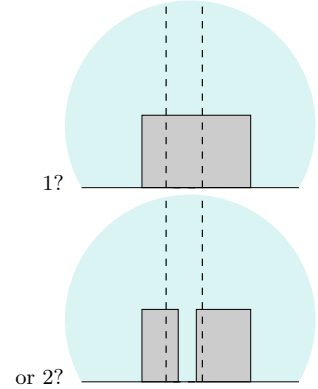


Figure 28.7. How many boxes are behind the tree?

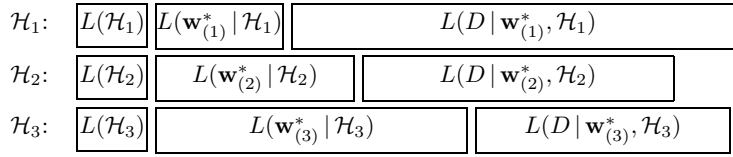


Figure 28.8. A popular view of model comparison by minimum description length. Each model \mathcal{H}_i communicates the data D by sending the identity of the model, sending the best-fit parameters of the model \mathbf{w}^* , then sending the data relative to those parameters. As we proceed to more complex models the length of the parameter message increases. On the other hand, the length of the data message decreases, because a complex model is able to fit the data better, making the residuals smaller. In this example the intermediate model \mathcal{H}_2 achieves the optimum trade-off between these two trends.

► 28.3 Minimum description length (MDL)

A complementary view of Bayesian model comparison is obtained by replacing probabilities of events by the lengths in bits of messages that communicate the events without loss to a receiver. Message lengths $L(\mathbf{x})$ correspond to a probabilistic model over events \mathbf{x} via the relations:

$$P(\mathbf{x}) = 2^{-L(\mathbf{x})}, \quad L(\mathbf{x}) = -\log_2 P(\mathbf{x}). \quad (28.15)$$

The MDL principle (Wallace and Boulton, 1968) states that one should prefer models that can communicate the data in the smallest number of bits. Consider a two-part message that states which model, \mathcal{H} , is to be used, and then communicates the data D within that model, to some pre-arranged precision δD . This produces a message of length $L(D, \mathcal{H}) = L(\mathcal{H}) + L(D | \mathcal{H})$. The lengths $L(\mathcal{H})$ for different \mathcal{H} define an implicit prior $P(\mathcal{H})$ over the alternative models. Similarly $L(D | \mathcal{H})$ corresponds to a density $P(D | \mathcal{H})$. Thus, a procedure for assigning message lengths can be mapped onto posterior probabilities:

$$L(D, \mathcal{H}) = -\log P(\mathcal{H}) - \log (P(D | \mathcal{H}) \delta D) \quad (28.16)$$

$$= -\log P(\mathcal{H} | D) + \text{const.} \quad (28.17)$$

In principle, then, MDL can always be interpreted as Bayesian model comparison and *vice versa*. However, this simple discussion has not addressed how one would actually evaluate the key data-dependent term $L(D | \mathcal{H})$, which corresponds to the evidence for \mathcal{H} . Often, this message is imagined as being subdivided into a parameter block and a data block (figure 28.8). Models with a small number of parameters have only a short parameter block but do not fit the data well, and so the data message (a list of large residuals) is long. As the number of parameters increases, the parameter block lengthens, and the data message becomes shorter. There is an optimum model complexity (\mathcal{H}_2 in the figure) for which the sum is minimized.

This picture glosses over some subtle issues. We have not specified the precision to which the parameters \mathbf{w} should be sent. This precision has an important effect (unlike the precision δD to which real-valued data D are sent, which, assuming δD is small relative to the noise level, just introduces an additive constant). As we decrease the precision to which \mathbf{w} is sent, the parameter message shortens, but the data message typically lengthens because the truncated parameters do not match the data so well. There is a non-trivial optimal precision. In simple Gaussian cases it is possible to solve for this optimal precision (Wallace and Freeman, 1987), and it is closely related to the posterior error bars on the parameters, \mathbf{A}^{-1} , where $\mathbf{A} = -\nabla \nabla \ln P(\mathbf{w} | D, \mathcal{H})$. It turns out that the optimal parameter message length is virtually identical to the log of the Occam factor in equation (28.10). (The random element involved in parameter truncation means that the encoding is slightly sub-optimal.)

With care, therefore, one can replicate Bayesian results in MDL terms. Although some of the earliest work on complex model comparison involved the MDL framework (Patrick and Wallace, 1982), MDL has no apparent advantages over the direct probabilistic approach.

MDL does have its uses as a pedagogical tool. The description length concept is useful for motivating prior probability distributions. Also, different ways of breaking down the task of communicating data using a model can give helpful insights into the modelling process, as will now be illustrated.

On-line learning and cross-validation.

In cases where the data consist of a sequence of points $D = \mathbf{t}^{(1)}, \mathbf{t}^{(2)}, \dots, \mathbf{t}^{(N)}$, the log evidence can be decomposed as a sum of ‘on-line’ predictive performances:

$$\begin{aligned} \log P(D | \mathcal{H}) &= \log P(\mathbf{t}^{(1)} | \mathcal{H}) + \log P(\mathbf{t}^{(2)} | \mathbf{t}^{(1)}, \mathcal{H}) \\ &+ \log P(\mathbf{t}^{(3)} | \mathbf{t}^{(1)}, \mathbf{t}^{(2)}, \mathcal{H}) + \dots + \log P(\mathbf{t}^{(N)} | \mathbf{t}^{(1)} \dots \mathbf{t}^{(N-1)}, \mathcal{H}). \end{aligned} \quad (28.18)$$

This decomposition can be used to explain the difference between the evidence and ‘leave-one-out cross-validation’ as measures of predictive ability. Cross-validation examines the average value of just the last term, $\log P(\mathbf{t}^{(N)} | \mathbf{t}^{(1)} \dots \mathbf{t}^{(N-1)}, \mathcal{H})$, under random re-orderings of the data. The evidence, on the other hand, sums up how well the model predicted all the data, starting from scratch.

The ‘bits-back’ encoding method.

Another MDL thought experiment (Hinton and van Camp, 1993) involves incorporating random bits into our message. The data are communicated using a parameter block and a data block. The parameter vector sent is a random sample from the posterior, $P(\mathbf{w} | D, \mathcal{H}) = P(D | \mathbf{w}, \mathcal{H})P(\mathbf{w} | \mathcal{H})/P(D | \mathcal{H})$. This sample \mathbf{w} is sent to an arbitrary small granularity $\delta\mathbf{w}$ using a message length $L(\mathbf{w} | \mathcal{H}) = -\log[P(\mathbf{w} | \mathcal{H})\delta\mathbf{w}]$. The data are encoded relative to \mathbf{w} with a message of length $L(D | \mathbf{w}, \mathcal{H}) = -\log[P(D | \mathbf{w}, \mathcal{H})\delta D]$. Once the data message has been received, the random bits used to generate the sample \mathbf{w} from the posterior can be deduced by the receiver. The number of bits so recovered is $-\log[P(\mathbf{w} | D, \mathcal{H})\delta\mathbf{w}]$. These recovered bits need not count towards the message length, since we might use some other optimally-encoded message as a random bit string, thereby communicating that message at the same time. The net description cost is therefore:

$$\begin{aligned} L(\mathbf{w} | \mathcal{H}) + L(D | \mathbf{w}, \mathcal{H}) - \text{‘Bits back’} &= -\log \frac{P(\mathbf{w} | \mathcal{H}) P(D | \mathbf{w}, \mathcal{H}) \delta D}{P(\mathbf{w} | D, \mathcal{H})} \\ &= -\log P(D | \mathcal{H}) - \log \delta D. \end{aligned} \quad (28.19)$$

Thus this thought experiment has yielded the optimal description length. Bits-back encoding has been turned into a practical compression method for data modelled with latent variable models by Frey (1998).

Further reading

Bayesian methods are introduced and contrasted with sampling-theory statistics in (Jaynes, 1983; Gull, 1988; Lored, 1990). The Bayesian Occam’s razor is demonstrated on model problems in (Gull, 1988; MacKay, 1992a). Useful textbooks are (Box and Tiao, 1973; Berger, 1985).

One debate worth understanding is the question of whether it’s permissible to use improper priors in Bayesian inference (Dawid *et al.*, 1996). If we want to do model comparison (as discussed in this chapter), it is essential to use proper priors – otherwise the evidences and the Occam factors are

meaningless. Only when one has no intention to do model comparison may it be safe to use improper priors, and even in such cases there are pitfalls, as Dawid *et al.* explain. I would agree with their advice to *always use proper priors*, tempered by an encouragement to be smart when making calculations, recognizing opportunities for approximation.

► 28.4 Exercises

Exercise 28.1.^[3] Random variables x come independently from a probability distribution $P(x)$. According to model \mathcal{H}_0 , $P(x)$ is a uniform distribution

$$P(x|\mathcal{H}_0) = \frac{1}{2} \quad x \in (-1, 1). \tag{28.20}$$

According to model \mathcal{H}_1 , $P(x)$ is a nonuniform distribution with an unknown parameter $m \in (-1, 1)$:

$$P(x|m, \mathcal{H}_1) = \frac{1}{2}(1 + mx) \quad x \in (-1, 1). \tag{28.21}$$

Given the data $D = \{0.3, 0.5, 0.7, 0.8, 0.9\}$, what is the evidence for \mathcal{H}_0 and \mathcal{H}_1 ?

Exercise 28.2.^[3] Datapoints (x, t) are believed to come from a straight line. The experimenter chooses x , and t is Gaussian-distributed about

$$y = w_0 + w_1 x \tag{28.22}$$

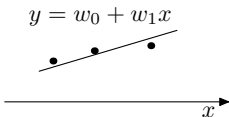
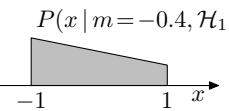
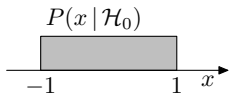
with variance σ_ν^2 . According to model \mathcal{H}_1 , the straight line is horizontal, so $w_1 = 0$. According to model \mathcal{H}_2 , w_1 is a parameter with prior distribution $\text{Normal}(0, 1)$. Both models assign a prior distribution $\text{Normal}(0, 1)$ to w_0 . Given the data set $D = \{(-8, 8), (-2, 10), (6, 11)\}$, and assuming the noise level is $\sigma_\nu = 1$, what is the evidence for each model?

Exercise 28.3.^[3] A six-sided die is rolled 30 times and the numbers of times each face came up were $\mathbf{F} = \{3, 3, 2, 2, 9, 11\}$. What is the probability that the die is a perfectly fair die (\mathcal{H}_0), assuming the alternative hypothesis \mathcal{H}_1 says that the die has a biased distribution \mathbf{p} , and the prior density for \mathbf{p} is uniform over the simplex $p_i \geq 0, \sum_i p_i = 1$?

Solve this problem two ways: exactly, using the helpful Dirichlet formulae (23.30, 23.31), and approximately, using Laplace's method. Notice that your choice of basis for the Laplace approximation is important. See MacKay (1998a) for discussion of this exercise.

Exercise 28.4.^[3] The influence of race on the imposition of the death penalty for murder in America has been much studied. The following three-way table classifies 326 cases in which the defendant was convicted of murder. The three variables are the defendant's race, the victim's race, and whether the defendant was sentenced to death. (Data from M. Radelet, 'Racial characteristics and imposition of the death penalty,' *American Sociological Review*, **46** (1981), pp.918-927.)

White defendant			Black defendant		
	Death penalty			Death penalty	
	Yes	No		Yes	No
White victim	19	132	White victim	11	52
Black victim	0	9	Black victim	6	97



It seems that the death penalty was applied much more often when the victim was white than when the victim was black. When the victim was white 14% of defendants got the death penalty, but when the victim was black 6% of defendants got the death penalty. [Incidentally, these data provide an example of a phenomenon known as *Simpson's paradox*: a higher fraction of white defendants are sentenced to death overall, but in cases involving black victims a higher fraction of black defendants are sentenced to death and in cases involving white victims a higher fraction of black defendants are sentenced to death.]

Quantify the evidence for the four alternative hypotheses shown in figure 28.9. I should mention that I don't believe any of these models is adequate: several additional variables are important in murder cases, such as whether the victim and murderer knew each other, whether the murder was premeditated, and whether the defendant had a prior criminal record; none of these variables is included in the table. So this is an academic exercise in model comparison rather than a serious study of racial bias in the state of Florida.

The hypotheses are shown as graphical models, with arrows showing dependencies between the variables v (victim race), m (murderer race), and d (whether death penalty given). Model \mathcal{H}_{00} has only one free parameter, the probability of receiving the death penalty; model \mathcal{H}_{11} has four such parameters, one for each state of the variables v and m . Assign uniform priors to these variables. How sensitive are the conclusions to the choice of prior?

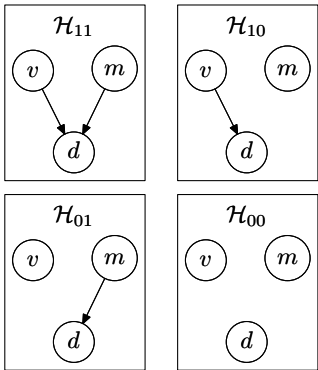


Figure 28.9. Four hypotheses concerning the dependence of the imposition of the death penalty d on the race of the victim v and the race of the convicted murderer m . \mathcal{H}_{01} , for example, asserts that the probability of receiving the death penalty does depend on the murderer's race, but not on the victim's.