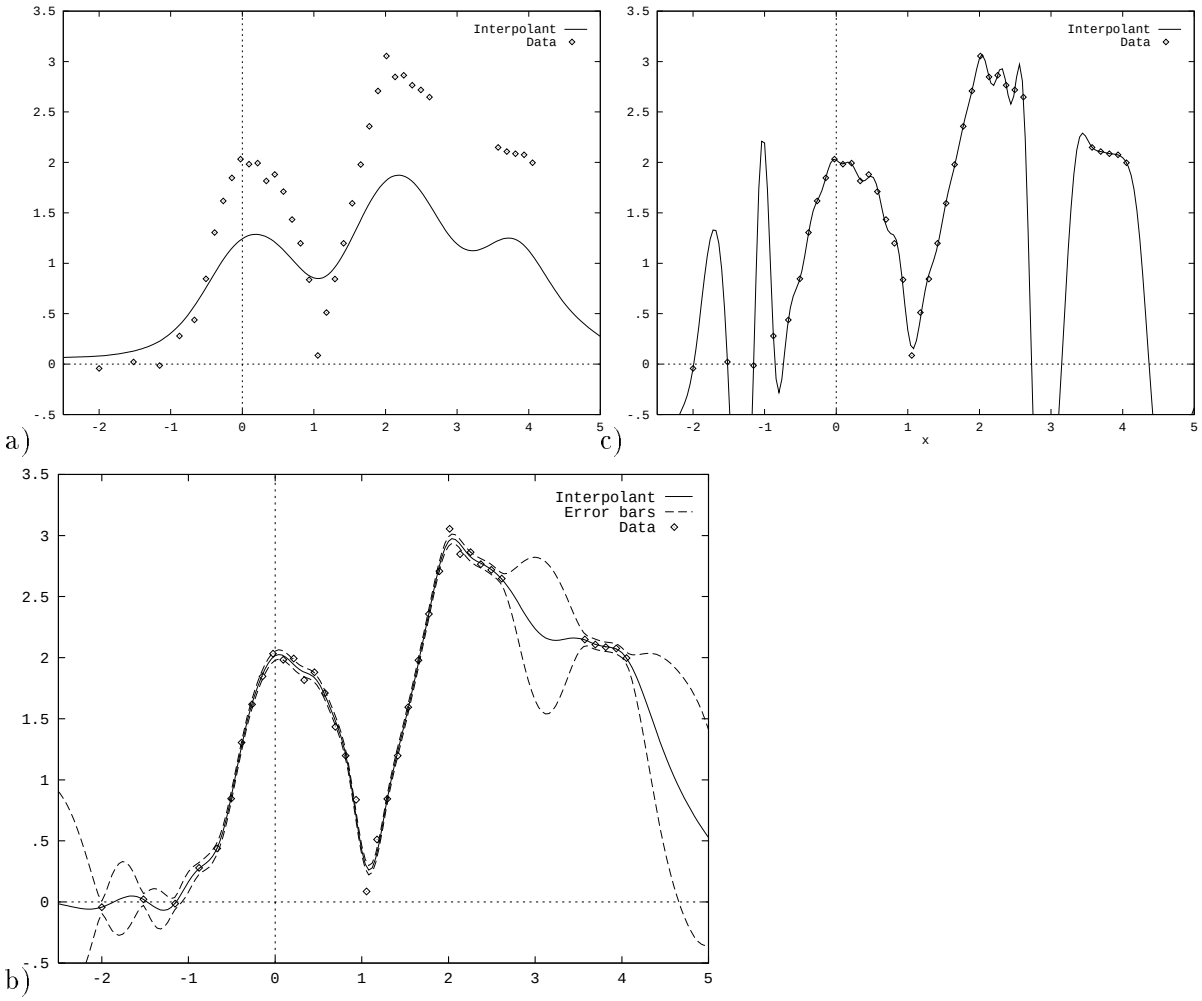

Figure 2.4: **How the best interpolant depends on α**

These figures introduce a data set, 'X,' which is interpolated with a variety of models in this chapter. Notice that the density of data points is not uniform on the x–axis. In the three figures the data set is interpolated using a radial basis function model with a basis of 60 equally spaced Cauchy functions, all with radius 0.2975. The regulariser is $E_W = \frac{1}{2}\sum w^2$, where w are the coefficients of the basis functions. Each figure shows the most probable interpolant for a different value of α: a) 6000; b) 2.5; c) 10^{-7}. Note at the extreme values how the data are oversmoothed and overfitted respectively. Assuming a flat prior, $\alpha = 2.5$ is the *most probable* value of α. In b), the most probable interpolant is displayed with its 1σ error bars, which represent how uncertain we are about the interpolant at each point, under the assumption that the interpolation model and the value of α are correct. Notice how the error bars increase in magnitude where the data are sparse. The error bars do not get bigger near the datapoint close to (1,0), because the radial basis function model does not expect sharp discontinuities; the error bars are obtained *assuming the model is correct,* so that point is interpreted as an improbable outlier.

Writing[7]

$$M(\mathbf{w}) = \alpha E_W + \beta E_D, \tag{2.13}$$

the posterior is

$$P(\mathbf{w}|D, \alpha, \beta, \mathcal{A}, \mathcal{R}, \mathcal{N}) = \frac{\exp(-M(\mathbf{w}))}{Z_M(\alpha, \beta)} \tag{2.14}$$

where $Z_M(\alpha, \beta) = \int d^k\mathbf{w} \exp(-M)$. We see that minimising the combined objective function M corresponds to finding the *most probable interpolant*, $\mathbf{w}_{\rm MP}$. Error bars on the best fit interpolant[8] can be obtained from the Hessian of M, $\mathbf{A} = \nabla\nabla M$, evaluated at $\mathbf{w}_{\rm MP}$.

This is the well known Bayesian view of regularisation [53, 72].

Bayesian methods provide far more than just an interpretation for regularisation. What we have described so far is just the first of three levels of inference. (The second level described in sections 1 and 2, 'model comparison,' splits into a second and a third level for this problem, because each interpolation model is made up of a continuum of sub–models with different values of α and β.) At the second level, Bayes allows us to objectively assign values to α and β, which are commonly unknown *a priori*. At the third, Bayes enables us to quantitatively rank alternative basis sets $\mathcal{A}$, alternative regularisers (priors) $\mathcal{R}$, and, in principle, alternative noise models $\mathcal{N}$. Furthermore, we can quantitatively compare interpolation under any model $\mathcal{H} = \{\mathcal{A}, \mathcal{N}, \mathcal{R}\}$ with other interpolation and learning models such as neural networks, if a similar Bayesian approach is applied to them. Neither the second nor the third level of inference can be successfully executed without Occam's razor.

The Bayesian theory of the second and third levels of inference has only recently been worked out [22], and this chapter's goal is to review that framework. Section 2.4 will describe the Bayesian method of inferring α and β; section 2.5 will describe Bayesian model comparison for the interpolation problem. Both these inference problems are solved by evaluating the appropriate evidence.

2.4 Selection of parameters α and β

Typically, α is not known *a priori*, and often β is also unknown. As α is varied, the properties of the best fit (most probable) interpolant vary. Assume that we are using a prior that encourages smoothness, and imagine that we interpolate at a very large value of α; then this will constrain the interpolant to be very smooth and flat, and it will not fit the data at all well (figure 2.4a). As α is decreased, the interpolant starts to fit the data better (figure 2.4b). If α is made even smaller, the interpolant oscillates wildly so as to overfit the noise in the data (figure 2.4c). The choice of the 'best' value of α is our first 'Occam's razor' problem: large values of α correspond to simple models which make constrained and precise predictions, saying 'the interpolant is expected to not have extreme curvature anywhere'; a tiny value of α corresponds to the more powerful and flexible model that says 'the interpolant could be anything at all, our prior belief in smoothness is very weak.' The task is to find a value of α which is small enough that the data are fitted but not so small that they are overfitted. For more severely ill–posed problems such as deconvolution, the precise value of the regularising parameter is increasingly important. Orthodox statistics has ways of assigning values to such parameters, based for example on misfit criteria, the

[7] The name M stands for 'misfit'; it will be demonstrated later that M is the natural measure of misfit, rather than $\chi^2_D = 2\beta E_D$.

[8] These error bars represent the uncertainty of the interpolant, and should not be confused with the typical scatter of noisy data points relative to the interpolant.

is assumed to have the form:

$$y(x) = \sum_{h=1}^{k} w_h \phi_h(x) \tag{2.7}$$

where the parameters w_h are to be inferred from the data. The data set is modelled as deviating from this mapping under some additive noise process $\mathcal{N}$:

$$t_m = y(x_m) + \nu_m \tag{2.8}$$

If ν is modelled as zero–mean gaussian noise with standard deviation σ_ν, then the probability of the data[4] given the parameters $\mathbf{w}$ is:

$$P(D\,|\mathbf{w},\beta,\mathcal{A},\mathcal{N}) = \frac{\exp(-\beta E_D(D|\mathbf{w},\mathcal{A}))}{Z_D(\beta)} \tag{2.9}$$

where $\beta = 1/\sigma_\nu^2$, $E_D = \sum_m \frac{1}{2}(y(x_m) - t_m)^2$, and $Z_D = (2\pi/\beta)^{N/2}$. $P(D\,|\mathbf{w},\beta,\mathcal{A},\mathcal{N})$ is called the likelihood. It is well known that finding the maximum likelihood parameters $\mathbf{w}_{\rm ML}$ may be an 'ill–posed' problem. That is, the $\mathbf{w}$ that minimises E_D is underdetermined and/or depends sensitively on the details of the noise in the data; the maximum likelihood interpolant in such cases oscillates wildly so as to fit the noise. Thus it is clear that to complete an interpolation model we need a prior $\mathcal{R}$ that expresses the sort of smoothness we expect the interpolant $y(x)$ to have. A model may have a prior of the form

$$P(y|\mathcal{R},\alpha) = \frac{\exp(-\alpha E_y(y|\mathcal{R}))}{Z_y(\alpha)} \tag{2.10}$$

where E_y might be for example the functional $E_y = \int y''(x)^2 dx$ (which is the regulariser for cubic spline interpolation[5]). The parameter α is a measure of how smooth $f(x)$ is expected to be. Such a prior can also be written as a prior on the parameters $\mathbf{w}$:

$$P(\mathbf{w}|\alpha,\mathcal{A},\mathcal{R}) = \frac{\exp(-\alpha E_W(\mathbf{w}|\mathcal{A},\mathcal{R}))}{Z_W(\alpha)} \tag{2.11}$$

where $Z_W = \int d^k\mathbf{w}\, \exp(-\alpha E_W)$. E_W (or E_y) is commonly referred to as a regularising function.

The interpolation model $\mathcal{H}$ is now complete, consisting of a choice of basis functions $\mathcal{A}$, a noise model $\mathcal{N}$ with parameter β, and a prior (regulariser) $\mathcal{R}$, with regularising constant α.

The first level of inference

If α and β are known, then the posterior probability of the parameters $\mathbf{w}$ is:[6]

$$P(\mathbf{w}|D,\alpha,\beta,\mathcal{A},\mathcal{R},\mathcal{N}) = \frac{P(D|\mathbf{w},\beta,\mathcal{A})P(\mathbf{w}|\alpha,\mathcal{A},\mathcal{R})}{P(D|\alpha,\beta,\mathcal{A},\mathcal{R},\mathcal{N})} \tag{2.12}$$

[4] Strictly, this probability should be written $P(\{t_m\}|\{x_m\},\mathbf{w},\beta,\mathcal{A},\mathcal{N})$, since these interpolation models do not predict the distribution of input variables $\{x_m\}$; this liberty of notation will be taken throughout this thesis.

[5] Strictly, this particular prior may be improper because a $y(x)$ of the form $w_1 x + w_0$ is not constrained by this prior.

[6] The regulariser $\alpha, \mathcal{R}$ has been omitted from the conditioning variables in the likelihood because the data distribution does not depend on the prior once $\mathbf{w}$ is known. Similarly the prior does not depend on $\beta, \mathcal{N}$.

Comments

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

- It is common for there to be degeneracies in models with many parameters, *i.e.* several equivalent parameters could be relabelled without affecting the likelihood. In these cases, the right hand side of equation (2.6) should be multiplied by the degeneracy of $\mathbf{w}_{\rm MP}$ to give the correct estimate of the evidence.

- 'Minimum description length' (MDL) methods are closely related to this Bayesian framework [55, 74, 75]. The log evidence $\log_2 P(D|\mathcal{H}_i)$ is the number of bits in the ideal shortest message that encodes the data D using model $\mathcal{H}_i$. Akaike's criterion [3] can be viewed as an approximation to MDL [58, 78]. Any implementation of MDL necessitates approximations in evaluating the length of the ideal shortest message. Although some of the earliest work on complex model comparison involved the MDL framework [50], MDL has no apparent advantages, and in my work I approximate the evidence directly.

- It should be emphasised that the Occam factor has nothing to do with how computationally complex it is to *use* a model. The evidence is a measure of *plausibility* of a model. How much CPU time it takes to use each model is certainly an interesting issue which might bias our decisions towards simpler models, but Bayes' rule does not address that. Choosing between models on the basis of how many function calls they need is an exercise in *decision theory*, which is not addressed in this chapter. Once the probabilities described above have been inferred, optimal actions can be chosen using standard decision theory with a suitable utility function.

2.3 The noisy interpolation problem

Bayesian interpolation through *noise–free* data has been studied by Skilling and Sibisi [60]. In this chapter I study the problem of interpolating through data where the dependent variables are assumed to be noisy (a task also known as 'regression,' 'curve–fitting,' 'signal estimation,' or, in the neural networks community, 'learning'). I am not examining the case where the independent variables are also noisy. This different and more difficult problem has been studied for the case of straight line–fitting by Gull [23].

Let us assume that the data set to be interpolated is a set of pairs $D = \{x_m, t_m\}$, where $m = 1 \ldots N$ is a label running over the pairs. For simplicity I will treat x and t as scalars, but the method generalises to the multidimensional case. To define a linear interpolation model, a set of k fixed basis functions[3] $\mathcal{A} = \{\phi_h(x)\}$ is chosen, and the interpolated function

[3] the case of *adaptive* basis functions, also known as feedforward neural networks, is examined in Chapter 3.

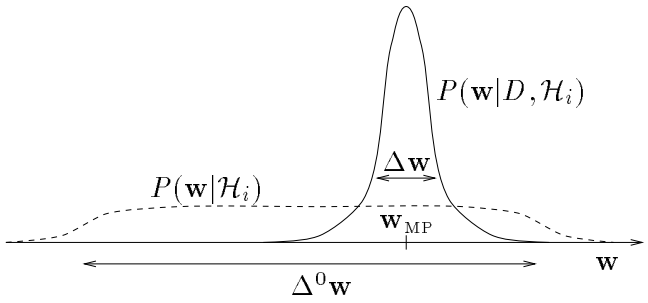

Figure 2.3: **The Occam factor**
This figure shows the quantities that determine the Occam factor for a hypothesis $\mathcal{H}_i$ having a single parameter $\mathbf{w}$. The prior distribution (dotted line) for the parameter has width $\Delta^0\mathbf{w}$. The posterior distribution (solid line) has a single peak at $\mathbf{w}_{\rm MP}$ with characteristic width $\Delta\mathbf{w}$. The Occam factor is $\frac{\Delta\mathbf{w}}{\Delta^0\mathbf{w}}$.

$\mathcal{H}_i$ thought possible before the data arrived (figure 2.3). Then $P(\mathbf{w}_{\rm MP}|\mathcal{H}_i) = \frac{1}{\Delta^0\mathbf{w}}$, and

$$\text{Occam factor} = \frac{\Delta\mathbf{w}}{\Delta^0\mathbf{w}},$$

i.e. **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 [21, 31]. The model $\mathcal{H}_i$ can be viewed as being composed of a certain number of equivalent submodels, of which only one survives when the data arrive. The Occam factor is the inverse of that number. The log of the Occam factor can be interpreted as the amount of information we gain about the model when the data arrive.

Typically, a complex model with many parameters, each of which is free to vary over a large range $\Delta^0\mathbf{w}$, will be penalised with a larger Occam factor than a simpler model. The Occam factor also provides a penalty for models which have to be finely tuned to fit the data; the Occam factor promotes models for which the required precision of the parameters $\Delta\mathbf{w}$ is coarse. The Occam factor is thus a measure of complexity of the model, but unlike the V–C dimension or algorithmic complexity, it relates to the complexity of the predictions that the model makes in data space; therefore it depends on the number of data points and other properties of the data set. Which model achieves the greatest evidence is determined by a trade–off between minimising this natural complexity measure and minimising the data misfit.

Occam factor for several parameters

If $\mathbf{w}$ is k-dimensional, and if the posterior is well approximated by a gaussian, the Occam factor is given by the determinant of the gaussian's covariance matrix:

$$\begin{aligned} P(D\,|\mathcal{H}_i) &\simeq \underbrace{P(D\,|\mathbf{w}_{\rm MP}, H_i)}_{\text{Best fit likelihood}}\ \underbrace{P(\mathbf{w}_{\rm MP}|\mathcal{H}_i)\,(2\pi)^{k/2}\det{}^{-\frac{1}{2}}\mathbf{A}}_{\text{Occam factor}} \\ \text{Evidence} &\simeq \text{Best fit likelihood} \qquad \text{Occam factor} \end{aligned} \tag{2.6}$$

where $\mathbf{A} = -\nabla\nabla \log P(\mathbf{w}|D,\mathcal{H}_i)$, the Hessian which we already evaluated when we calculated the error bars on $\mathbf{w}_{\rm MP}$. As the amount of data collected, N, increases, this gaussian approximation is expected to become increasingly accurate on account of the central limit theorem [73]. For the linear interpolation models discussed in this chapter, this gaussian expression is exact for any N.

ancient stone circles (for which some people have proposed extremely elaborate theories!), discuss a practical method of assigning relative prior probabilities to alternative models by evaluating the lengths of the computer programs that decode data previously encoded under each model [50]. This procedure introduces a second sort of Occam's razor into the inference, namely a *prior* bias against complex models. However, we will not include such prior biases here; we will address only the data's preference for the alternative models, *i.e.* the evidence, and the Occam's razor that it embodies. In the limit of large quantities of data this objective Occam's razor will always be the more important of the two.

A modern Bayesian approach to priors

It should be pointed out that the emphasis of this modern Bayesian approach is not on the inclusion of priors into inference. There is not one significant 'subjective prior' in this entire chapter. (The reader who is interested to see problems where significant subjective priors do arise is directed to [23, 63].) The emphasis is that consistent degrees of preference for alternative hypotheses are represented by probabilities, and relative preferences for models are assigned by evaluating those probabilities. Historically, Bayesian analysis has been accompanied by methods to work out the 'right' prior $P(\mathbf{w}|\mathcal{H})$ for a problem, for example, the principles of insufficient reason and maximum entropy. The modern Bayesian however does not take a fundamentalist attitude to assigning the 'right' priors — many different priors can be tried; each particular prior corresponds to a different hypothesis about the way the world is. We can compare these alternative hypotheses in the light of the data by evaluating the evidence. This is the way in which alternative regularisers are compared, for example. If we try one model and obtain awful predictions, we have *learnt* something. 'A failure of Bayesian prediction is an opportunity to learn' [30], and we are able to come back to the same data set with new models, using new priors for example.

Evaluating the evidence

Let us now explicitly study the evidence to gain insight into how the Bayesian Occam's razor works. The evidence is the normalising constant for equation (2.1):

$$P(D\,|\mathcal{H}_i) = \int P(D|\mathbf{w},\mathcal{H}_i)P(\mathbf{w}|\mathcal{H}_i)\,d\mathbf{w} \tag{2.4}$$

For many problems, including interpolation, it is common for the posterior $P(\mathbf{w}|D,\mathcal{H}_i) \propto P(D|\mathbf{w},\mathcal{H}_i)P(\mathbf{w}|\mathcal{H}_i)$ to have a strong peak at the most probable parameters $\mathbf{w}_{\rm MP}$ (figure 2.3). Then the evidence can be approximated 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, $\Delta\mathbf{w}$:

$$\begin{aligned} P(D\,|\mathcal{H}_i) \;&\simeq\; \underbrace{P(D\,|\mathbf{w}_{\rm MP},\mathcal{H}_i)}\;\underbrace{P(\mathbf{w}_{\rm MP}|\mathcal{H}_i)\Delta\mathbf{w}} \\ \text{Evidence} \;&\simeq\; \text{Best fit likelihood}\;\;\text{Occam factor} \end{aligned} \tag{2.5}$$

Thus the evidence is found by taking the best fit likelihood that the model can achieve and multiplying it by an 'Occam factor' [21], which is a term with magnitude less than one that penalises $\mathcal{H}_i$ for having the parameter $\mathbf{w}$.

Interpretation of the Occam factor

The quantity $\Delta\mathbf{w}$ is the posterior uncertainty in $\mathbf{w}$. Imagine for simplicity that the prior $P(\mathbf{w}|\mathcal{H}_i)$ is uniform on some large interval $\Delta^0\mathbf{w}$, representing the range of values of $\mathbf{w}$ that

the Hessian $\mathbf{A} = -\nabla\nabla \log P(\mathbf{w}|D, \mathcal{H}_i)$ and Taylor–expanding the log posterior with $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_{\rm MP}$,

$$P(\mathbf{w}|D, \mathcal{H}_i) \simeq P(\mathbf{w}_{\rm MP}|D, \mathcal{H}_i) \exp\left(-\tfrac{1}{2}\Delta\mathbf{w}^{\rm T}\mathbf{A}\Delta\mathbf{w}\right) \tag{2.2}$$

we see that the posterior can be locally approximated as a gaussian with covariance matrix (error bars) $\mathbf{A}^{-1}$.[2]

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:

$$P(\mathcal{H}_i|D) \propto P(D|\mathcal{H}_i)P(\mathcal{H}_i) \tag{2.3}$$

Notice that the data–dependent term $P(D|\mathcal{H}_i)$ is the evidence for $\mathcal{H}_i$, which appeared as the normalising constant in (2.1). The second term, $P(\mathcal{H}_i)$, is a 'subjective' prior over our hypothesis space which expresses how plausible we thought the alternative models were before the data arrived. We will see later that this subjective part of the inference will typically be overwhelmed by the objective term, the evidence. Assuming that we have no reason to assign strongly differing priors $P(\mathcal{H}_i)$ to the alternative models, **models $\mathcal{H}_i$ are ranked by evaluating the evidence.** Equation (2.3) has not been normalised because in the data modelling process we may develop new models after the data have arrived (figure 2.1), when an inadequacy of the first models is detected, for example. So we do not start with a completely defined hypothesis space. Inference is open–ended: we continually seek more probable models to account for the data we gather. New models are compared with previous models by evaluating the evidence for them.

The key concept of this chapter is this: to assign a preference to 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; whether our data modelling task is a regression problem, a classification problem, or a density estimation problem, the evidence is the Bayesian's transportable quantity for comparing alternative models. In all these cases the evidence naturally embodies Occam's razor; we will examine how this works shortly.

Of course, the evidence is not the whole story if we have good reason to assign unequal priors to the alternative models $\mathcal{H}$. (To only use the evidence for model comparison is equivalent to using maximum likelihood for parameter estimation.) The classic example is the 'Sure Thing' hypothesis, © E.T Jaynes, which is the hypothesis that the data set will be D, the precise data set that actually occurred; the evidence for the Sure Thing hypothesis is huge. But Sure Thing belongs to an immense class of similar hypotheses which should all be assigned correspondingly tiny prior probabilities; so the posterior probability for Sure Thing is negligible alongside any sensible model. Models like Sure Thing are rarely seriously proposed in real life, but if such models are developed then clearly we need to think about precisely what priors are appropriate. Patrick and Wallace, studying the geometry of

[2] Whether this approximation is a good one or not will depend on the problem we are solving. For the interpolation models discussed in this chapter, there is only a single maximum in the posterior distribution, and the gaussian approximation is exact. For more general statistical models we still expect the posterior to be dominated by locally gaussian peaks on account of the central limit theorem [73]. Multiple maxima which arise in more complex models complicate the analysis, but Bayesian methods can still successfully be applied [25, 41, 46].

Bayesian methods to achieve the state of the art in image reconstruction. The same approach to regularisation has also been developed in part by Szeliski [70]. Bayesian model comparison is also discussed by Bretthorst [9], who has used Bayesian methods to push back the limits of NMR signal detection. The same Bayesian theory underlies the unsupervised classification system, AutoClass [25]. The fact that Bayesian model comparison embodies Occam's razor has been rediscovered by Kashyap in the context of modelling time series [33]; his paper includes a thorough discussion of how Bayesian model comparison is different from orthodox 'Hypothesis testing.' One of the earliest applications of these sophisticated Bayesian methods of model comparison to real data is by Patrick and Wallace [50]; in this fascinating paper, competing models accounting for megalithic stone circle geometry are compared within the description length framework, which is equivalent to Bayes.

As the quantities of data collected throughout science and engineering continue to increase, and the computational power and techniques available to model that data also multiply, I believe Bayesian methods will prove an ever more important tool for refining our modelling abilities. I hope that this review will help to introduce these techniques to the 'neural' modelling community. Chapter 3 will demonstrate how these techniques can be fruitfully applied to backpropagation neural networks. Chapter 4 will show how this framework relates to the task of selecting where next to gather data so as to gain maximal information about our models.

2.2 The evidence and the Occam factor

Let us write down Bayes' rule for the two levels of inference described above, so as to see explicitly how Bayesian model comparison works. Each model $\mathcal{H}_i$ ($\mathcal{H}$ stands for 'hypothesis') is assumed to have a vector of parameters $\mathbf{w}$. A model is defined by its functional form and two probability distributions: a 'prior' distribution $P(\mathbf{w}|\mathcal{H}_i)$ which states what values the model's parameters might plausibly take; and the predictions $P(D|\mathbf{w},\mathcal{H}_i)$ that the model makes about the data D when its parameters have a particular value $\mathbf{w}$. Note that models with the same parameterisation but different priors over the parameters are therefore defined to be different models.

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

$$P(\mathbf{w}|D,\mathcal{H}_i) = \frac{P(D|\mathbf{w},\mathcal{H}_i)P(\mathbf{w}|\mathcal{H}_i)}{P(D|\mathcal{H}_i)} \tag{2.1}$$

   In words:

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

   The normalising constant $P(D|\mathcal{H}_i)$ is commonly ignored, since it is irrelevant to the first level of inference, *i.e.* the choice of $\mathbf{w}$; but it will be important in the second level of inference, and we name it the **evidence** for $\mathcal{H}_i$. It is common to use gradient–based methods to find the maximum of the posterior, which defines the most probable value for the parameters, $\mathbf{w}_{\text{MP}}$; it is then common to summarise the posterior distribution by the value of $\mathbf{w}_{\text{MP}}$, and error bars on these best fit parameters. The error bars are obtained from the curvature of the posterior; writing

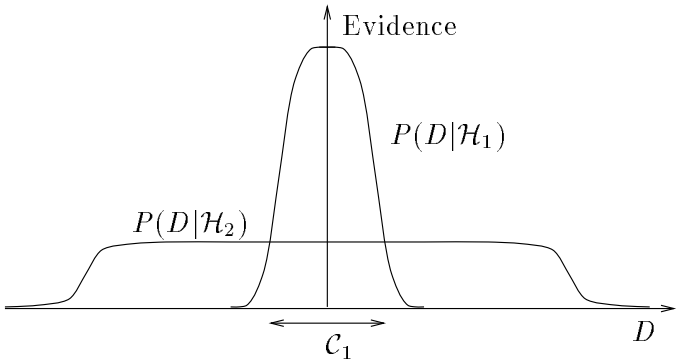

Figure 2.2: **Why Bayes embodies Occam's razor**
This figure gives the basic intuition for why complex models are penalised. The horizontal axis represents the space of possible data sets D. Bayes rule rewards models in proportion to how much they *predicted* the data that occurred. These predictions are quantified by a normalised probability distribution on D. In this paper, 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 $\mathcal{C}_1$ as strongly as $\mathcal{H}_1$. Assume that equal prior probabilities have been assigned to the two models. Then if the data set falls in region $\mathcal{C}_1$, the *less powerful* model $\mathcal{H}_1$ will be the *more probable* 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–parameterised models which generalise poorly. 'Occam's razor' is the principle that states that unnecessarily complex models should not be preferred to simpler ones. Bayesian methods automatically and quantitatively embody Occam's razor [21, 31], without the introduction of ad hoc penalty terms. Complex models are automatically self–penalising under Bayes' rule. Figure 2.2 gives the basic intuition for why this should be expected; the rest of this chapter will explore this property in depth.

Bayesian methods were first laid out in depth by the Cambridge geophysicist Sir Harold Jeffreys [31]. The logical basis for the Bayesian use of probabilities as measures of plausibility was subsequently established by Cox [12], who proved that consistent inference in a closed hypothesis space can be mapped onto probabilities. For a general review of Bayesian philosophy the reader is encouraged to read the excellent papers by Jaynes and Loredo [30, 38]. Since Jeffreys the emphasis of most Bayesian probability theory has been 'to formally utilize prior information' [7], *i.e.* to perform inference in a way that makes explicit the prior knowledge and ignorance that we have, which orthodox methods omit. However, Jeffreys' work also laid the foundation for Bayesian model comparison, which does not involve an emphasis on prior information. Jeffreys applied this theory to simple model comparison problems in geophysics, for example testing whether a single additional parameter is justified by the data. Since the 1960s, Jeffreys' model comparison methods have been applied and extended in the economics literature [78]. Only recently has this aspect of Bayesian analysis been further developed and applied to more complex problems in other fields.

This chapter will review Bayesian model comparison, 'regularisation,' and noise estimation, by studying the problem of interpolating noisy data. The Bayesian framework I will describe for these tasks is due to Gull and Skilling [21, 22, 24, 60, 64], who have used

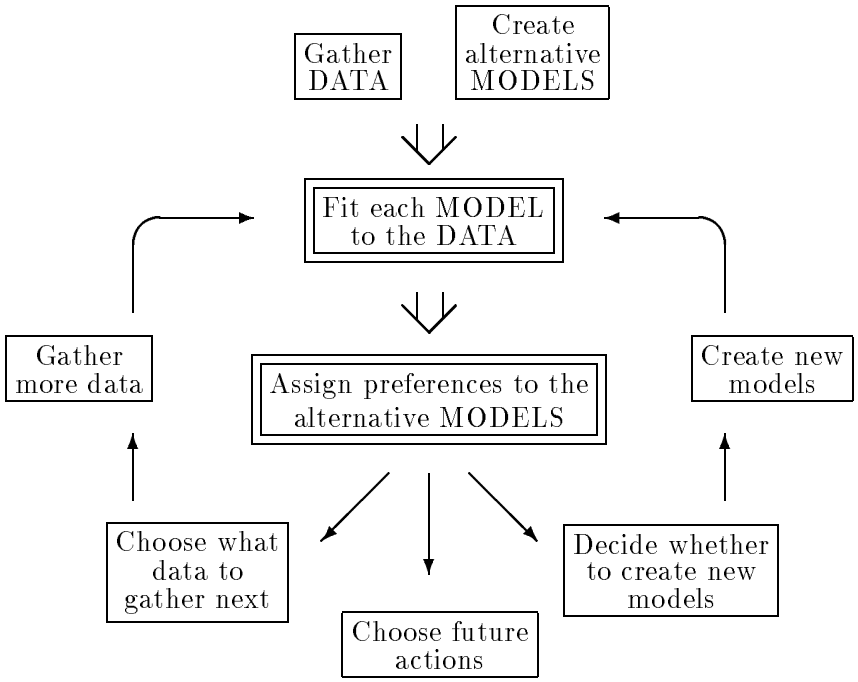

Figure 2.1: **Where Bayesian inference fits into the data modelling process.** This figure illustrates an abstraction of the part of the scientific process, in which data is collected and modelled. In particular, this figure applies to pattern classification, learning, interpolation, etc.. The two double–framed boxes denote the two steps which involve *inference*. It is only in those two steps that Bayes' rule 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. Bayes may be used to find the most probable parameter values, and error bars on those parameters. The result of applying Bayes to this problem is often little different from the result of using orthodox statistics.
The second inference task, model comparison in the light of the data, is where Bayes is in a class of its own. This second inference problem requires a quantitative Occam's razor to penalise over–complex models. Bayes can assign objective preferences to the alternative models in a way that automatically embodies Occam's razor.

feedforward neural networks. At the first level of inference, we take each model individually and find the best fit interpolant for that model. At the second level of inference we want to rank the alternative models and state for our particular data set that, for example, 'splines are probably the best interpolation model,' or 'if the interpolant is modelled as a polynomial, it should probably be a cubic.'

Bayesian methods are able consistently and quantitatively to solve both these inference tasks. There is a popular myth that states that Bayesian methods only differ from orthodox (also known as 'frequentist' or 'sampling theory') statistical methods by the inclusion of subjective priors which are arbitrary and difficult to assign, and 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 Bayes performs the second level of inference. It is here that Bayesian methods are totally different from orthodox methods. Indeed, when regression and density estimation are discussed in most statistics texts (for example [19]), the task of model comparison is virtually ignored; no general orthodox method exists for solving this problem.

Model comparison is a difficult task because it is not possible simply to choose the model

Chapter 2

Bayesian Interpolation

Abstract

Although Bayesian analysis has been in use since Laplace, the Bayesian method of *model–comparison* has only recently been developed in depth. In this chapter, the Bayesian approach to regularisation and model–comparison is demonstrated by studying the inference problem of interpolating noisy data. The concepts and methods described are quite general and can be applied to many other data modelling problems.

Regularising constants are set by examining their posterior probability distribution. Alternative regularisers (priors) and alternative basis sets are objectively compared by evaluating the *evidence* for them. 'Occam's razor' is automatically embodied by this process.

The way in which Bayes infers the values of regularising constants and noise levels has an elegant interpretation in terms of the effective number of parameters determined by the data set. This framework is due to Gull and Skilling.

2.1 Data modelling and Occam's razor

In science, a central task is to develop and compare models to account for the data that are gathered. In particular this is true in the problems of learning, pattern classification, interpolation and clustering. Two levels of **inference** are involved in the task of data modelling (figure 2.1). At the first level of inference, we assume that one of the models that we invented is true, and we fit that model to the data. Typically a model includes some free parameters; fitting the model to the data involves inferring what values those parameters should probably take, given the data. The results of this inference are often summarised by the most probable parameter values and error bars on those parameters. This 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.[1]

For example, consider the task of interpolating a noisy data set. The data set could be interpolated using a splines model, using radial basis functions, using polynomials, or using

[0] Chapter 2 of PhD thesis 'Bayesian Methods for Adaptive Models' by David MacKay, California Institute of Technology, submitted December 10 1991.

[1] 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 minimise the expectation of a 'loss function.' This chapter concerns inference alone and no loss functions or utilities are involved.