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 an interpolation model, a set of $k$ fixed basis functions[1] $\mathcal{A} = \{\phi_h(x)\}$ is chosen, and the interpolated function is assumed to have the form:

$$y(x) = \sum_{h=1}^{k} w_h \phi_h(x) \tag{7}$$

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:

$$t_m = y(x_m) + \nu_m \tag{8}$$

If $\nu$ is modelled as zero–mean gaussian noise with standard deviation $\sigma_\nu$, then the probability of the data[2] given the parameters $\mathbf{w}$ is:

$$P(D\,|\mathbf{w}, \beta, \mathcal{A}) = \frac{\exp -\beta E_D(D|\mathbf{w}, \mathcal{A})}{Z_D(\beta)} \tag{9}$$

where $\beta = 1/\sigma_\nu^2$, $E_D = \sum_m (y(x_m) - t_m)^2$, and $Z_D = (2\pi/\beta)^{N/2}$. $P(D\,|\mathbf{w}, \beta, \mathcal{A})$ 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. To complete the interpolation model we need a prior $\mathcal{R}$ that expresses the sort of smoothness we expect the interpolant $y(x)$ to have. We may have a prior of the form

$$P(y|\mathcal{R}, \alpha) = \frac{\exp -\alpha E_y(y|\mathcal{R})}{Z_y(\alpha)} \tag{10}$$

where $E_y$ might be for example the functional $E_y = \int y''(x)^2 dx$ (which is the regulariser for cubic spline interpolation[3]). The parameter $\alpha$ 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 $\mathbf{w}$:

$$P(\mathbf{w}|\mathcal{A}, \mathcal{R}, \alpha) = \frac{\exp -\alpha E_W(\mathbf{w}|\mathcal{A}, \mathcal{R})}{Z_W(\alpha)} \tag{11}$$

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 is now complete, consisting of a choice of basis functions $\mathcal{A}$, a noise model with parameter $\beta$, and a prior (regulariser) $\mathcal{R}$, with regularising constant $\alpha$.

[1] the case of *adaptive* basis functions, also known as feedforward neural networks, is examined in a companion paper.

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

[3] Strictly, this particular prior may be improper because an $f$ of the form $w_1 x + w_0$ is not constrained by this prior.

## The first level of inference

If $\alpha$ and $\beta$ are known, then the posterior probability of the parameters $\mathbf{w}$ is:[4]

$$P(\mathbf{w}|D, \alpha, \beta, \mathcal{A}, \mathcal{R}) = \frac{P(D|\mathbf{w}, \beta, \mathcal{A}) P(\mathbf{w}|\alpha, \mathcal{A}, \mathcal{R})}{P(D|\alpha, \beta, \mathcal{A}, \mathcal{R})} \tag{12}$$

Writing[5]

$$M(\mathbf{w}) = \alpha E_W + \beta E_D, \tag{13}$$

the posterior is

$$P(\mathbf{w}|D, \alpha, \beta, \mathcal{A}, \mathcal{R}) = \frac{\exp -M(\mathbf{w})}{Z_M(\alpha, \beta)} \tag{14}$$

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 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 [15, 23].

Bayes can do a lot more than just provide an interpretation for regularisation. What we have described so far is just the first of three levels of inference. (The second level of model comparison described in sections 1 and 2 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 $\alpha$ and $\beta$.) At the second level, Bayes allows us to objectively assign values to $\alpha$ and $\beta$, which are commonly unknown *a priori*. At the third, Bayes enables us to quantitatively rank alternative basis sets $\mathcal{A}$, and regularisers (priors) $\mathcal{R}$ (and, in principle, alternative noise models). Furthermore, we can quantitatively compare interpolation under any model $\mathcal{A}, \mathcal{R}$ 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 succesfully executed without Occam's razor.

The Bayesian theory of the second and third levels of inference has only recently been worked out, and this paper's goal is to review that framework. Section 4 will describe the Bayesian method of choosing $\alpha$ and $\beta$; section 5 will describe Bayesian model comparison for the interpolation problem. Both these inference problems are solved by evaluating the appropriate evidence.

# 4 Selection of parameters $\alpha$ and $\beta$

Typically, $\alpha$ is not known *a priori*, and often $\beta$ is also unknown. As $\alpha$ is varied, the properties of the best fit (most

[4] The regulariser $\mathcal{R}$ 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 $\beta$.

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

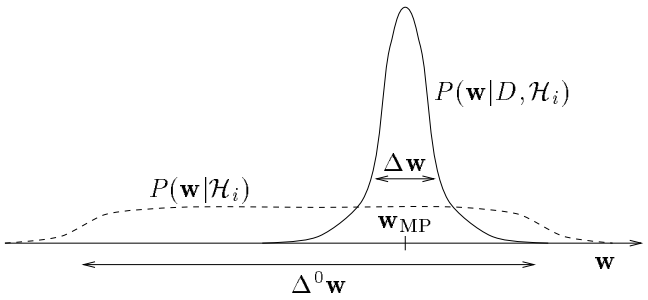


Figure 3: **The Occam factor**
This figure shows the quantities that determine the Occam factor for a hypothesis $\mathcal{H}_i$ having a single parameter $\mathbf{w}$. The prior distribution (dotted line) for the parameter has width $\Delta^0\mathbf{w}$. The posterior distribution (solid line) has a single peak at $\mathbf{w}_{\rm MP}$ with characteristic width $\Delta\mathbf{w}$. The Occam factor is $\frac{\Delta\mathbf{w}}{\Delta^0\mathbf{w}}$.

is the normalising constant for equation (1):

$$P(D\,|\mathcal{H}_i) = \int P(D|\mathbf{w},\mathcal{H}_i)P(\mathbf{w}|\mathcal{H}_i)d\mathbf{w} \qquad (4)$$

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 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, $\Delta\mathbf{w}$:

$$\begin{array}{rcll} P(D\,|\mathcal{H}_i) & \simeq & \underbrace{P(D\,|\mathbf{w}_{\rm MP},\mathcal{H}_i)} & \underbrace{P(\mathbf{w}_{\rm MP}|\mathcal{H}_i)\,\Delta\mathbf{w}} \\ \text{Evidence} & \simeq & \text{Best fit likelihood} & \text{Occam factor} \end{array} \qquad (5)$$

Thus the evidence is found by taking the best fit likelihood that the model can achieve and multiplying it by an 'Occam factor' [5], 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

$\Delta\mathbf{w}$ 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 $\mathcal{H}_i$ thought possible before the data arrived. 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 arrives [5, 11]. The log of the Occam factor can be interpreted as the amount of information we gain about the model $\mathcal{H}$ when the data arrives.

In summary, 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. 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:

$$\begin{array}{rcll} P(D\,|\mathcal{H}_i) & \simeq & \underbrace{P(D\,|\mathbf{w}_{\rm MP},H_i)} & \underbrace{P(\mathbf{w}_{\rm MP}|\mathcal{H}_i)\,(2\pi)^{k/2}\det{}^{-\frac{1}{2}}\mathbf{A}} \\ \text{Evidence} & \simeq & \text{Best fit likelihood} & \text{Occam factor} \end{array} \qquad (6)$$

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 $\mathbf{w}_{\rm MP}$.

### 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 $\mathbf{A}$. 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 relabeled without affecting the likelihood. In these cases, the right hand side of equation (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. 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 criteria are an approximation to MDL [16, 24, 25]. Any implementation of MDL necessitates approximations in evaluating the length of the ideal shortest message. I can see no advantage in MDL, and recommend that the evidence should be approximated directly.

## 3 The noisy interpolation problem

Bayesian interpolation through noise–free data has been studied by Skilling and Sibisi [17]. In this paper I study the case where the dependent variables are assumed to be noisy. I am not however examining the case where the independent variables are noisy too. This different and more difficult problem has been studied for the case of straight line–fitting by Gull [7].

## 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.

1. **Model fitting.** Assuming that one model $\mathcal{H}_i$ is true, 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:

$$P(\mathbf{w}|D,\mathcal{H}_i) = \frac{P(D|\mathbf{w},\mathcal{H}_i)P(\mathbf{w}|\mathcal{H}_i)}{P(D|\mathcal{H}_i)} \qquad (1)$$

   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 methods to find the maximum of the posterior, which defines the most probable value for the parameters, $\mathbf{w}_{\rm MP}$; error bars on these best fit parameters can be obtained from the curvature of the posterior. Writing the Hessian $\mathbf{A} = -\nabla\nabla \log P(\mathbf{w}|D,\mathcal{H}_i)$ and Taylor–expanding the log posterior with $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_{\rm MP}$,

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

   we see that the posterior may be locally approximated as a gaussian with covariance matrix (error bars) $\mathbf{A}^{-1}$. For the interpolation models discussed in this paper, there is only a single maximum in the posterior distribution, and the gaussian approximation is exact; but this is of course not the case for a general problem. Multiple maxima complicate the analysis, but Bayesian methods can still successfully be applied [13, 14].

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:

$$P(\mathcal{H}_i|D) \propto P(D|\mathcal{H}_i)P(\mathcal{H}_i) \qquad (3)$$

   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 (1). Assuming that we have no reason to assign strongly differing priors $P(\mathcal{H}_i)$ to the alternative hypotheses, **hypotheses $\mathcal{H}_i$ are ranked by evaluating the evidence.** Equation (3) has not been normalised because in the scientific process we may develop new models after the data have arrived (figure 1), when a failure of the first models occurs, for example. So we do not start with a completely defined space of hypotheses. 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 evidence is the Bayesian's transportable quantity for comparing alternative hypotheses.

The key concept of this paper is this: to assign a preference to alternative models $\mathcal{H}_i$, a Bayesian evaluates the evidence $P(D|\mathcal{H}_i)$.

Of course, the evidence is not the whole story if we have good reason to assign unequal priors to the alternative hypotheses $\mathcal{H}$. (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 occured; 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. Clearly if models such as this one are developed then we will need to think about precisely what priors are appropriate. However models like Sure Thing are rarely seriously proposed in real life.

### 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, as is widely held. There is not one significant 'subjective prior' in this entire paper. (If you are interested to see problems where subjective priors do arise see [7, 20].) The emphasis is that degrees of preference for alternative hypotheses are represented by probabilities, and relative preferences for hypotheses are assigned by evaluating those probabilities. Historically Bayesian analysis has been accompanied by methods to work out the 'right' prior for a problem. The modern Bayesian does not take a fundamentalist attitude to assigning the 'right' priors — many different priors can be tried; any particular prior corresponds to a 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 hypothesis and obtain awful predictions, we have *learnt* something. A failure of Bayesian prediction is an opportunity to learn, and we are able to come back to the same data set with new hypotheses, 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

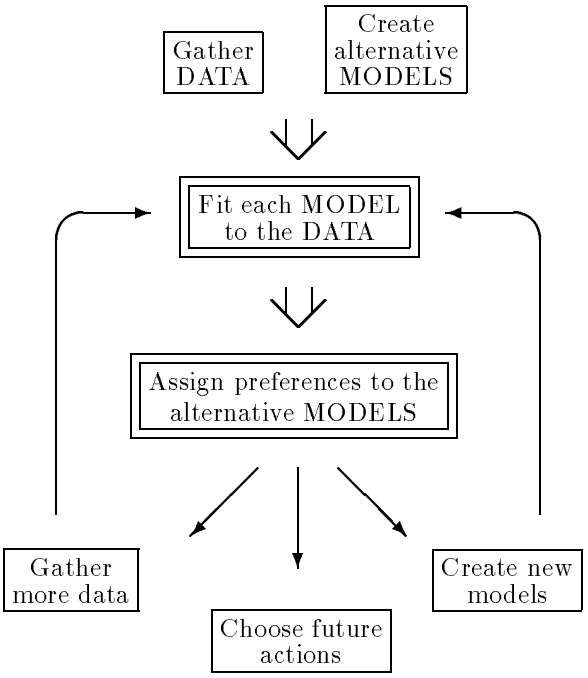


Figure 1: **Where Bayesian inference fits into science, in particular pattern classification, learning, interpolation, etc.**
This figure illustrates an abstraction of a central part of the scientific process, and many other processes involving the collecting and modelling of data. The two double–framed boxes denote the two steps which involve *inference*. It is only in those two steps that Bayes can be used. Bayes does not tell you how to invent hypotheses, 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 hypotheses in a way that automatically embodies Occam's razor.

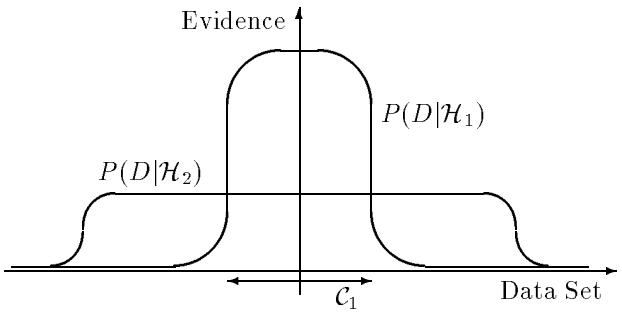


Figure 2: **Why Bayes embodies Occam's razor**
This figure gives the basic intuition for why complex hypotheses are penalised. The horizontal axis represents the space of possible data sets $D$. Bayes rule rewards hypotheses in proportion to how much they *predicted* the data that occurred. These predictions are quantified by a normalised probability distribution on $D$. In this paper, this probability of the data given model $\mathcal{H}_i$ is called the evidence, $P(D|\mathcal{H}_i)$.
A simple hypothesis $\mathcal{H}_1$ makes only a limited range of predictions, shown by $P(D|\mathcal{H}_1)$; a more powerful hypothesis $\mathcal{H}_2$, 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 hypotheses. Then if the data set falls in region $\mathcal{C}_1$, *the less powerful model $\mathcal{H}_1$ will be the more probable hypothesis.*

been 'to formally utilize prior information' [1], *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. Only recently has this aspect of Bayesian analysis been developed and applied to real world problems.

This paper 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 [5, 6, 8, 17, 18], who have used 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 [22]. Bayesian model comparison is also discussed by Bretthorst [2], who has used Bayesian methods to push back the limits of NMR signal detection.

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. A companion paper [13] will demonstrate how these techniques can be applied to backpropagation neural networks.

# Bayesian Interpolation

David J.C. MacKay
Computation and Neural Systems
California Institute of Technology 139–74
Pasadena CA 91125
mackay@hope.caltech.edu

May 27, 1991
Submitted to *Neural Computation* as a review paper

## 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 paper, 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 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 framework.

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.

## 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 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. We wish to compare the models in the light of the data, and assign some sort of preference or ranking to the alternatives.

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 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 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, model comparison is a task virtually ignored in most statistics texts, and 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 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. 'Occam's razor' is the principle that unnecessarily complex models should not be preferred to simpler ones. Bayesian methods automatically and quantitatively embody Occam's razor [5], without the introduction of ad hoc penalty terms. Complex hypotheses are automatically self–penalising under Bayes' rule. Figure 2 gives the basic intuition for why this should be expected.

Bayesian methods were first laid out in depth by Jeffreys [11]. For a general review of Bayesian philosophy see the excellent papers by Loredo and Jaynes [10, 12]. Since Jeffreys the emphasis of most Bayesian probability theory has