Bayesian Interpolation

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

May 18, 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 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

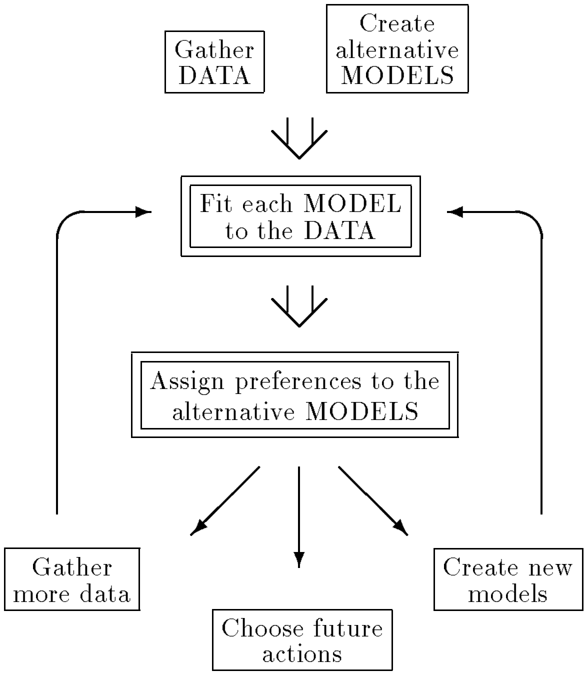

Gather DATA
Create alternative MODELS
Fit each MODEL to the DATA
Assign preferences to the alternative MODELS
Gather more data
Choose future actions
Create new models

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.

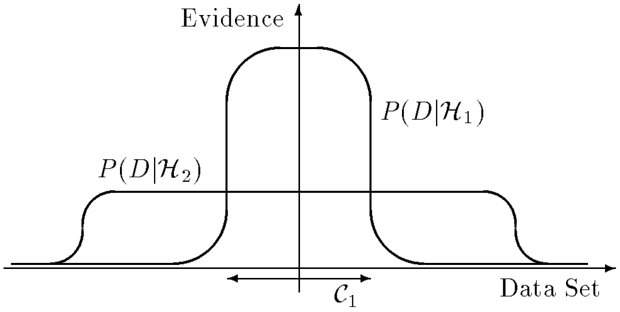


Figure 2: **Why Bayes embodies Occam's razor**
This figure gives the basic intuition for why complex hypotheses are penalised. The horizontal axis represents the space of possible data sets D. Bayes rule rewards hypotheses in proportion to how much they *predicted* the data that occurred. These predictions 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.*

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.

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:

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

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, $\mathbf{w}_{\text{MP}}$; 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 $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_{\text{MP}}$,

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

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:

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

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.

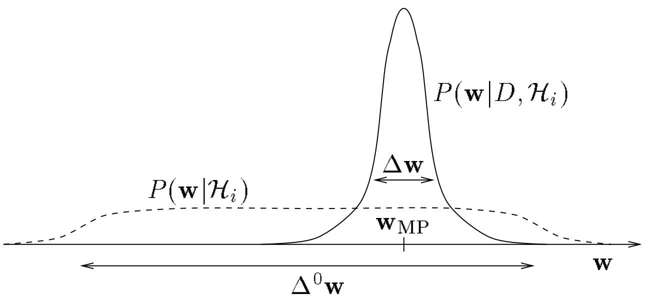
Evaluating the evidence

Let us now explicitly study the evidence to gain insight into how the Bayesian Occam's razor works. The evidence is the normalising constant for equation (1):

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

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 $\mathbf{w}_{\text{MP}}$ (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, $\Delta\mathbf{w}$:

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



$P(\mathbf{w}|D,\mathcal{H}_i)$
$\Delta\mathbf{w}$
$P(\mathbf{w}|\mathcal{H}_i)$
$\mathbf{w}_{\mathrm{MP}}$
$\mathbf{w}$
$\Delta^0\mathbf{w}$

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 $\mathbf{w}_{\mathrm{MP}}$ with characteristic width $\Delta\mathbf{w}$. The Occam factor is $\frac{\Delta\mathbf{w}}{\Delta^0\mathbf{w}}$.

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 $P(\mathbf{w}_{\mathrm{MP}}|\mathcal{H}_i) = \frac{1}{\Delta^0\mathbf{w}}$, 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:

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

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

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 $\mathbf{w}_{\mathrm{MP}}$ 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].

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:

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

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:

$$t_m = y(x_m) + \nu_m \qquad (8)$$

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

If ν is modelled as zero–mean gaussian noise with standard deviation σ_ν, 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 $\mathbf{w}_{\mathrm{ML}}$ 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 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 $\mathbf{w}$:

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

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 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:[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*, $\mathbf{w}_{\mathrm{MP}}$. Error bars on the best fit interpolant can be obtained from the hessian of M, $\mathbf{A} = \nabla\nabla M$, evaluated at $\mathbf{w}_{\mathrm{MP}}$.

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 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 $\mathcal{A}$, 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 α and β; 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 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 like the splines prior defined earlier, 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 4a). As α is decreased, the interpolant starts to fit the data better (figure 4b). If α is made even smaller, the interpolant oscillates wildly so as to overfit the noise in the data (figure 4c). The choice of the 'best' value of α is our first 'Occam's razor' problem: large values of α correspond to simple hypotheses 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 hypothesis 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 and cross–validation. Gull has demonstrated why the popular use of misfit criteria is incorrect and how Bayes sets these param-

[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.

[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 β.

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

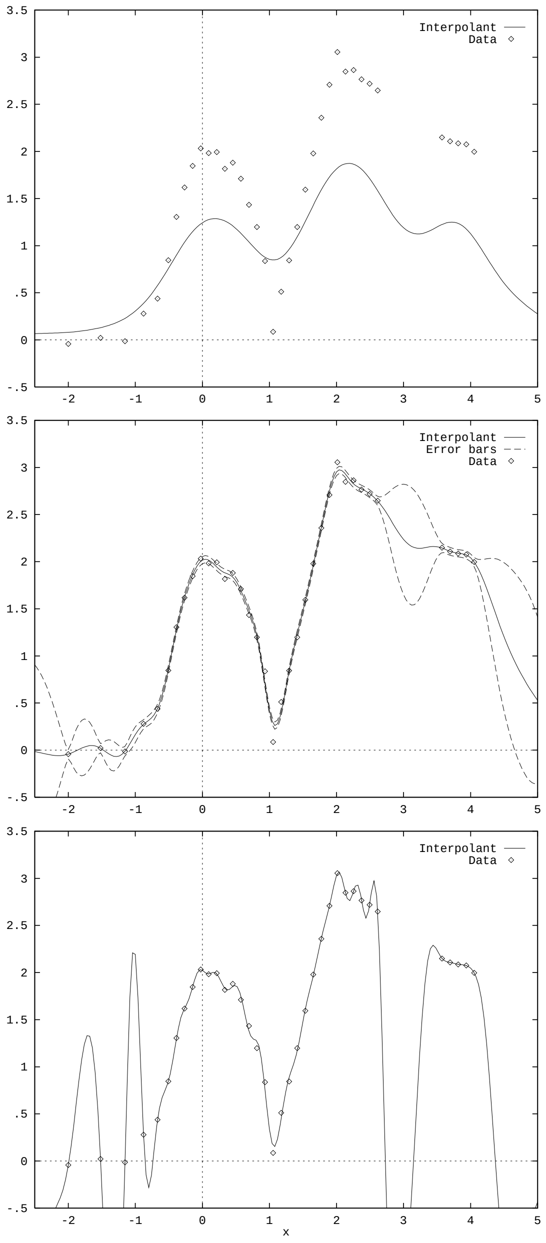


Figure 4: **How the best interpolant depends on α** These figures introduce a data set, 'X,' which is interpolated with a variety of models in this paper. Notice that the density of data points is not uniform on the x–axis. In the three figures the data set is interpolated using a radial basis function 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 is oversmoothed and overfitted respectively. $\alpha = 2.5$ 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 is sparse. The error bars do not include the datapoint close to (1,0), because the radial basis function model does not expect sharp discontinuities, so that point is interpreted as an improbable outlier.

eters [6]. Cross–validation may be an unreliable technique unless large quantities of data are available. See section 6 and [13] however for further discussion of cross–validation. I will explain the Bayesian method of setting α and β after first reviewing some statistics of misfit.

Misfit, χ^2, and the effect of parameter measurements

For N gaussian variables with mean μ and standard deviation σ, the statistic $\chi^2 = \sum (x-\mu)^2/\sigma^2$ is a measure of misfit. If μ is known *a priori*, χ^2 has expectation $N \pm \sqrt{N}$. However, if μ is fitted from the data by setting $\mu = \bar{x}$, we 'use up a degree of freedom', and χ^2 has expectation $N - 1$. In the second case μ is a 'well–measured parameter' (measured by the data). When a parameter is determined from the data in this way it is unavoidable that the parameter fits some of the noise as well. That is why the expectation of χ^2 is reduced by one. This is the basis of the distinction between the σ_N and σ_{N-1} buttons on your calculator. It is common for this distinction to be ignored, but in cases such as interpolation where the number of free parameters is similar to the number of data points, it is essential to find and make the analogous distinction. It will be demonstrated that the Bayesian choice of both α and β are most simply expressed in terms of the effective number of well–measured parameters, γ, to be derived below.

Misfit criteria are 'principles' which set parameters like α and β by requiring that χ^2 should have a particular value. The discrepancy principle requires $\chi^2 = N$. Another principle requires $\chi^2 = N - k$, where k is the number of free parameters. We will find that an intuitive misfit criterion arises for the optimal value of β; on the other hand, the Bayesian choice of α is unrelated to the misfit.

Bayesian choice of α and β

To infer from the data what value α and β should have,[6] Bayesians evaluate the posterior probability distribution:

$$P(\alpha,\beta|D,\mathcal{A},\mathcal{R}) = \frac{P(D|\alpha,\beta,\mathcal{A},\mathcal{R})P(\alpha,\beta)}{P(D|\mathcal{A},\mathcal{R})} \qquad (15)$$

The data dependent term $P(D|\alpha,\beta,\mathcal{A},\mathcal{R})$ has already appeared earlier as the normalising constant in equation (12), and it is called the evidence for α and β. Similarly the normalising constant of (15) is called the evidence for $\mathcal{A},\mathcal{R}$, and it will turn up later when we compare alternative models $\mathcal{A},\mathcal{R}$ in the light of the data.

If $P(\alpha,\beta)$ is a flat prior, the evidence is the function that we use to assign a preference to alternative values of α and

[6] Note that it is not satisfactory to simply maximise the likelihood over $\mathbf{w}$, α and β; the likelihood has a skew peak such that the maximum likelihood value for the parameters is not in the same place as most of the posterior probability. To get a feeling for this here is a more familiar problem: examine the posterior probability for the parameters of a gaussian (μ,σ) given N samples: the maximum likelihood value for σ is σ_N, but the most probable value for σ is σ_{N-1}.

β. It is given by

$$P(D|\alpha,\beta,\mathcal{A},\mathcal{R}) = \frac{Z_M(\alpha,\beta)}{Z_W(\alpha)Z_D(\beta)} \tag{16}$$

where the Z's have been defined earlier. Occam's razor is implicit in this formula: if α is small, the large freedom in the range of possible values of $\mathbf{w}$ is automatically penalised by the consequent large value of Z_W; models that fit the data well achieve a large value of Z_M; a model that has to be very finely tuned for it to fit the data is penalised by a smaller value of Z_M. The optimum value of α achieves a compromise between fitting the data well and being too powerful a model.

Now to assign a preference to (α,β), our computational task is to evaluate the three integrals Z_M, Z_W and Z_D. We will come back to this task in a moment.

But that sounds like determining your prior after the data have arrived!

This is an aside which can be omitted on a first reading. When I first heard the preceding explanation of Bayesian regularisation I was discontent because it seemed that the prior is being chosen from an ensemble of possible priors *after* the data have arrived. To be precise, as described above, the most probable prior (most probable value of α) is selected; then that prior alone is used to infer what the interpolant might be. This is not how Bayes would have us infer what the interpolant is. It is the combined ensemble of priors that define our prior, and we should integrate over this ensemble when we do inference. Let us work out what happens if we follow this proper approach. The preceding method of using only the most probable prior will emerge as a good approximation.

Let us examine the true posterior $P(\mathbf{w}|D,\mathcal{A},\mathcal{R})$, obtained by integrating over α and β:

$$P(\mathbf{w}|D,\mathcal{A},\mathcal{R}) = \int P(\mathbf{w}|D,\alpha,\beta,\mathcal{A},\mathcal{R})P(\alpha,\beta|D,\mathcal{A},\mathcal{R})\,d\alpha\,d\beta \tag{17}$$

In words, the posterior probability over $\mathbf{w}$ can be written as a linear combination of the posteriors for all values of α,β. Each posterior density is weighted by the probability of α,β given the data, which appeared in (15). This means that if $P(\alpha,\beta|D,\mathcal{A},\mathcal{R})$ has a single peak at $\hat{\alpha},\hat{\beta}$, then the true posterior $P(\mathbf{w}|D,\mathcal{A},\mathcal{R})$ will be dominated by the density $P(\mathbf{w}|D,\hat{\alpha},\hat{\beta},\mathcal{A},\mathcal{R})$. As long as the properties of the posterior $P(\mathbf{w}|D,\alpha,\beta,\mathcal{A},\mathcal{R})$ do not change rapidly with α,β near $\hat{\alpha},\hat{\beta}$ and the peak in $P(\alpha,\beta|D,\mathcal{A},\mathcal{R})$ is strong, we are justified in using the approximation:

$$P(\mathbf{w}|D,\mathcal{A},\mathcal{R}) \simeq P(\mathbf{w}|D,\hat{\alpha},\hat{\beta},\mathcal{A},\mathcal{R}) \tag{18}$$

Evaluating the evidence

Let us return to our train of thought at equation (16). To evaluate the evidence for α,β, we want to find the integrals Z_M, Z_W and Z_D. Typically the most difficult integral to evaluate is Z_M.

$$Z_M(\alpha,\beta) = \int d^k\mathbf{w}\ \exp -M(\mathbf{w},\alpha,\beta)$$

If the regulariser $\mathcal{R}$ is a quadratic functional (and the favourites are), then E_D and E_W are quadratic functions of $\mathbf{w}$, and we can evaluate Z_M exactly. Letting $\nabla\nabla E_W = \mathbf{C}$ and $\nabla\nabla E_D = \mathbf{B}$ then using $\mathbf{A} = \alpha\mathbf{C} + \beta\mathbf{B}$, we have:

$$M = M(\mathbf{w}_{\mathrm{MP}}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\mathrm{MP}})^{\mathrm{T}}\mathbf{A}(\mathbf{w} - \mathbf{w}_{\mathrm{MP}})$$

where $\mathbf{w}_{\mathrm{MP}} = \mathbf{A}^{-1}\mathbf{B}\mathbf{w}_{\mathrm{ML}}$. This means that Z_M is the gaussian integral:

$$Z_M = e^{-M_{\mathrm{MP}}}(2\pi)^{k/2}\det{}^{-\frac{1}{2}}\mathbf{A} \tag{19}$$

In many cases where the regulariser is not quadratic (for example, entropy–based), this gaussian approximation is still servicable [6]. Thus we can write the log evidence for α and β as:

$$\log P(D|\alpha,\beta,\mathcal{A},\mathcal{R}) = -\alpha E_W^{MP} - \beta E_D^{MP} - \frac{1}{2}\log\det\mathbf{A} - \log Z_W(\alpha) - \log Z_D(\beta) + \frac{k}{2}\log 2\pi \tag{20}$$

The term βE_D^{MP} represents the misfit of the interpolant to the data. The three terms $-\alpha E_W^{MP} - \frac{1}{2}\log\det\mathbf{A} - \log Z_W(\alpha)$ constitute the 'Occam factor' penalising over–powerful values of α: the ratio $(2\pi)^{k/2}\det^{-\frac{1}{2}}\mathbf{A}/Z_W(\alpha)$ is the ratio of the posterior accessible volume in parameter space to the prior accessible volume. Figure 5a illustrates the behaviour of these various terms as a function of α for the same radial basis function model as illustrated in figure 4.

Now we could just proceed to evaluate the evidence numerically as a function of α and β, but a more deep and fruitful understanding of this problem is possible.

Properties of the evidence maximum

The maximum over α,β of $P(D|\alpha,\beta,\mathcal{A},\mathcal{R}) = \frac{Z_M(\alpha,\beta)}{Z_W(\alpha)Z_D(\beta)}$ has some remarkable properties which give deeper insight into this Bayesian approach. The results of this section are useful both numerically and intuitively.

Following Gull [6], we transform to the basis in which the Hessian of E_W is the identity, $\nabla\nabla E_W = \mathbf{I}$. This transformation is simple in the case of quadratic E_W: rotate into the eigenvector basis of $\mathbf{C}$ and stretch the axes so that the quadratic form E_W becomes homogeneous. This is the natural basis for the prior. I will continue to refer to the parameter vector in this basis as $\mathbf{w}$. Using $\nabla\nabla M = \mathbf{A}$ and $\nabla\nabla E_D = \mathbf{B}$ as above, we differentiate the log evidence with respect to α and β so as to find the condition that is satisfied at the maximum. The log evidence, from (20), is:

$$\log P(D|\alpha,\beta,\mathcal{A},\mathcal{R}) = -\alpha E_W^{MP} - \beta E_D^{MP} - \frac{1}{2}\log\det\mathbf{A} + \frac{N}{2}\log\beta + \frac{k}{2}\log\alpha + \frac{k}{2}\log 2\pi \tag{21}$$

First, differentiating with respect to α, we need to evaluate $\frac{d}{d\alpha}\log\det\mathbf{A}$. Using $\mathbf{A} = \alpha\mathbf{I} + \beta\mathbf{B}$,

$$\begin{aligned}\frac{d}{d\alpha}\log\det\mathbf{A} &= \mathrm{Trace}\left(\mathbf{A}^{-1}\frac{d\mathbf{A}}{d\alpha}\right)\\ &= \mathrm{Trace}\mathbf{A}^{-1}\mathbf{I} = \mathrm{Trace}\mathbf{A}^{-1}\end{aligned}$$

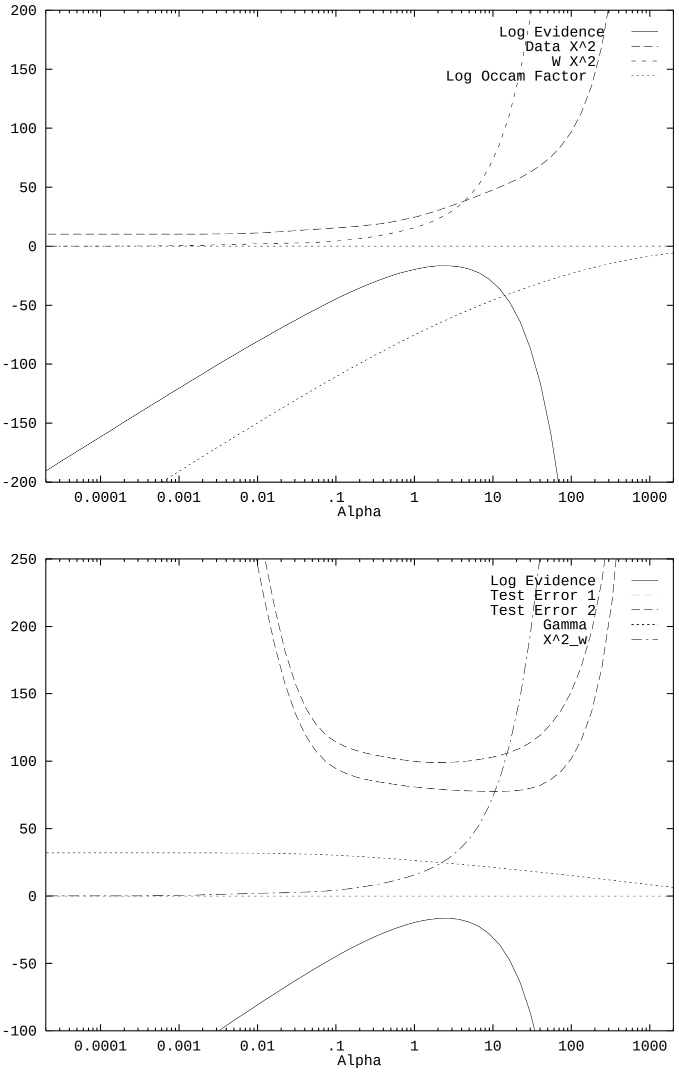


Figure 5: **Choosing α**
a) **The evidence as a function of α:** Using the same radial basis function model as in figure 4, this graph shows the log evidence as a function of α, and shows the functions which make up the log evidence, namely the data misfit $\chi^2_D = 2\beta E_D$, the weight penalty term $\chi^2_W = 2\alpha E_W$, and the log of the Occam factor $(2\pi)^{k/2}\det^{-\frac{1}{2}}\mathbf{A}/Z_W(\alpha)$.
b) **Criteria for optimising α:** This graph shows the log evidence as a function of α, and the functions whose intersection locates the evidence maximum: the number of good parameter measurements γ, and χ^2_W. Also shown is the test error (rescaled) on two test sets; finding the test error minimum is an alternative criterion for setting α. Both test sets were more than twice as large in size as the interpolated data set. Note how the point at which $\chi^2_W = \gamma$ is clear and unambiguous, which cannot be said for the minima of the test energies. The evidence gives α a confidence interval of $[1.3, 5.0]$. The test error minima are more widely distributed because of finite sample noise.

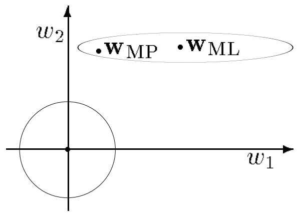


Figure 6: **Good and bad parameter measurements**
w_1 and w_2 are the components in parameter space in two directions parallel to eigenvectors of the data matrix $\mathbf{B}$. The circle represents the characteristic prior distribution for $\mathbf{w}$. The ellipse represents a characteristic contour of the likelihood, centred on the maximum likelihood solution $\mathbf{w}_{\rm ML}$. $\mathbf{w}_{\rm MP}$ represents the most probable parameter vector. w_1 is a direction in which λ_1 is small compared to α, *i.e.* the data have no strong preference about the value of w_1; w_1 is a poorly measured parameter, and the term $\frac{\lambda_1}{\lambda_1+\alpha}$ is close to zero. w_2 is a direction in which λ_1 is large; w_2 is well determined by the data, and the term $\frac{\lambda_2}{\lambda_2+\alpha}$ is close to one.

This result is exact if E_W and E_D are quadratic. Otherwise this result is an approximation, omitting terms in $\partial\mathbf{B}/\partial\alpha$. Now, differentiating (21) and setting the derivative to zero, we obtain the following condition for the most probable value of α:

$$2\alpha E_W^{MP} = k - \alpha \text{Trace}\mathbf{A}^{-1} \tag{22}$$

The quantity on the left is the dimensionless measure of the amount of structure introduced into the parameters by the data, *i.e.* how much the fitted parameters differ from their null value. It can be interpreted as the χ^2 of the parameters, since it is equal to $\chi^2_W = \sum w_i^2/\sigma^2_W$, with $\alpha = 1/\sigma^2_W$.

The quantity on the right of (22) is called the number of good parameter measurements, γ, and has value between 0 and k. It can be written in terms of the eigenvalues of $\beta\mathbf{B}$, λ_a. The eigenvalues of $\mathbf{A}$ are $\lambda_a + \alpha$, so we have:

$$\gamma = k - \alpha\,\text{Trace}\mathbf{A}^{-1} = k - \sum_{a=1}^{k} \frac{\alpha}{\lambda_a + \alpha} = \sum_{a=1}^{k} \frac{\lambda_a}{\lambda_a + \alpha} \tag{23}$$

Each eigenvalue λ_a measures how strongly one parameter is determined by the data. α measures how strongly the parameters are determined by the prior. The term $\lambda_a/(\lambda_a + \alpha)$ is a number between 0 and 1 which measures the strength of the data relative to the prior (figure 6). A direction in parameter space for which λ_a is small compared to α does not contribute to the number of good parameter measurements. γ is thus a measure of the effective number of parameters which are well determined by the data. As $\alpha/\beta \rightarrow 0$, γ increases from 0 to k.

This concept is not only important for locating the optimum value of α: it is only the γ good parameter measurements which are expected to contribute to the reduction of the data misfit that occurs when a model is fitted to noisy data. In the process of fitting $\mathbf{w}$ to the data, it is unavoidable that some fitting of the model to noise will occur, because some components of the noise are indistinguishable from real data. Typically, one unit (χ^2) of noise will be fitted for every well–determined parameter. Poorly determined parameters are determined by the regulariser only, so they do not reduce χ^2_D in this way. We will now examine how this concept enters into the Bayesian choice of β.

Recall that the expectation of the χ^2 misfit between the true interpolant and the data is N. We do not know the true interpolant, and the only misfit measure we have access to is the χ^2 of the data, $\chi^2_D = 2\beta E_D$. The 'discrepancy principle' of orthodox statistics states that the model parameters should be adjusted so as to make $\chi^2_D = N$. Other orthodox approaches would suggest that we should estimate the noise level so as to set $\chi^2_D = N - k$, where k is the number of free parameters. Let us find out the opinion of Bayes' rule on this matter.

We differentiate the log evidence (21) with respect to β and obtain:

$$2\beta E_D = N - \gamma \tag{24}$$

Thus the most probable noise estimate, $\hat{\beta}$, does not satisfy $\chi^2_D = N$ or $\chi^2_D = N - k$; rather, $\chi^2_D = N - \gamma$. Thus the Bayesian estimate of noise level naturally takes into account the fact that the parameters which have been determined by the data inevitably suppress some of the noise in the data, while the poorly determined parameters do not. Note that the value of χ^2_D only enters into the determination of β: misfit criteria have no role in the Bayesian choice of α [6].

In summary, at the optimum value of α and β, $\chi^2_W = \gamma$, $\chi^2_D = N - \gamma$. Notice that this implies that the total misfit $M = \alpha E_W + \beta E_D$ satisfies the simple equation $2M = N$.

The Bayesian choice of α is illustrated by figure 4b. Figure 5b illustrates the functions involved with the Bayesian choice of α, and compares them with the 'test error' approach. Demonstration of the Bayesian choice of β is omitted, since it is straightforward; β is fixed to its true value for the demonstrations in this paper.

These results generalise to the case where there are two or more separate regularisers with independent regularising constants $\alpha_1, \alpha_2 \ldots$ [6]. In this case, each regulariser has a number of well–measured parameters γ_i associated with it. Multiple regularisers will be used in the companion paper on neural networks.

Finding the evidence maximum with a head on approach would involve evaluating $\det \mathbf{A}$ while searching over α, β; the above results (22,24) enable us to speed up this search (for example by the use of re–estimation formulae like $\alpha := \gamma/2E_W$) and replace the evaluation of $\det \mathbf{A}$ by the evaluation of $\text{Trace}\mathbf{A}^{-1}$. For large dimensional problems where this task is demanding, Skilling has developed methods for estimating $\text{Trace}\mathbf{A}^{-1}$ statistically [21].

5 Bayesian model comparison

To rank alternative basis sets $\mathcal{A}$ and regularisers (priors) $\mathcal{R}$ in the light of the data, we examine the posterior probabilities:

$$P(\mathcal{A}, \mathcal{R}|D) \propto P(D|\mathcal{A}, \mathcal{R}) P(\mathcal{A}, \mathcal{R}) \tag{25}$$

The data–dependent term, the evidence for $\mathcal{A}, \mathcal{R}$, appeared earlier as the normalising constant in (15), and it is evaluated by integrating the evidence for (α, β):

$$P(D|\mathcal{A}, \mathcal{R}) = \int P(D|\mathcal{A}, \mathcal{R}, \alpha, \beta) P(\alpha, \beta)\, d\alpha\, d\beta \tag{26}$$

Assuming that we have no reason to assign strongly differing priors $P(\mathcal{A}, \mathcal{R})$, alternative $\mathcal{A}, \mathcal{R}$ are ranked just by examining the evidence. The evidence can also be compared with the evidence found by an equivalent Bayesian analysis of other learning and interpolation models so as to allow the data to assign a preference to the alternative models. Notice as pointed out earlier that this modern Bayesian framework includes no emphasis on defining the 'right' prior $\mathcal{R}$ with which we ought to interpolate. Rather, we invent as many priors (regularisers) as we want, and allow the data to tell us which prior is most probable.

Evaluating the evidence for $\mathcal{A}, \mathcal{R}$

As α and β vary, a single evidence maximum is obtained, at $\hat{\alpha}, \hat{\beta}$ (at least for quadratic E_D and E_W). The evidence maximum is usually well approximated[7] by a separable gaussian, and differentiating (21) twice we obtain gaussian error bars for $\log \alpha$ and $\log \beta$:

$$(\Delta \log \alpha)^2 \simeq 2/\gamma$$
$$(\Delta \log \beta)^2 \simeq 2/(N - \gamma)$$

Putting these error bars into (26), we obtain the evidence.[8]

$$P(D|\mathcal{A}, \mathcal{R}) \simeq P(D|\hat{\alpha}, \hat{\beta}, \mathcal{A}, \mathcal{R}) P(\hat{\alpha}, \hat{\beta})\, 2\pi\, \Delta\log \alpha\, \Delta\log \beta \tag{27}$$

How is the prior $P(\hat{\alpha}, \hat{\beta})$ assigned? This is the first time in this paper that we have met one of the infamous 'subjective priors', which are supposed to plague Bayesian methods. Here are some answers to this question. (a) Any other method of assigning a preference to alternatives must implicitly assign such priors. Bayesians adopt the healthy attitude of not sweeping them under the carpet. (b) With some thought, reasonable values can usually be assigned to subjective priors, and the degree of reasonable subjectivity in these assignments can be quantified. For example, a reasonable prior on an unknown standard deviation states that σ is unknown over a range of (3 ± 2) orders of magnitude. This prior contributes a subjectivity of about ± a factor of 2 to the value of the evidence. This degree of subjectivity is often negligible compared to the evidence differences. (c) In the noisy interpolation example, all models considered include the free parameters α and β. So in this paper I do not need to assign a value to $P(\hat{\alpha}, \hat{\beta})$; I assume that it is a flat prior which cancels out when we compare alternative interpolation models.

[7] Condition for this approximation: in the spectrum of eigenvalues of $\beta\mathbf{B}$, the number of eigenvalues within e–fold of α must be $O(1)$.

[8] There are analytic methods for performing such integrals over β [2].

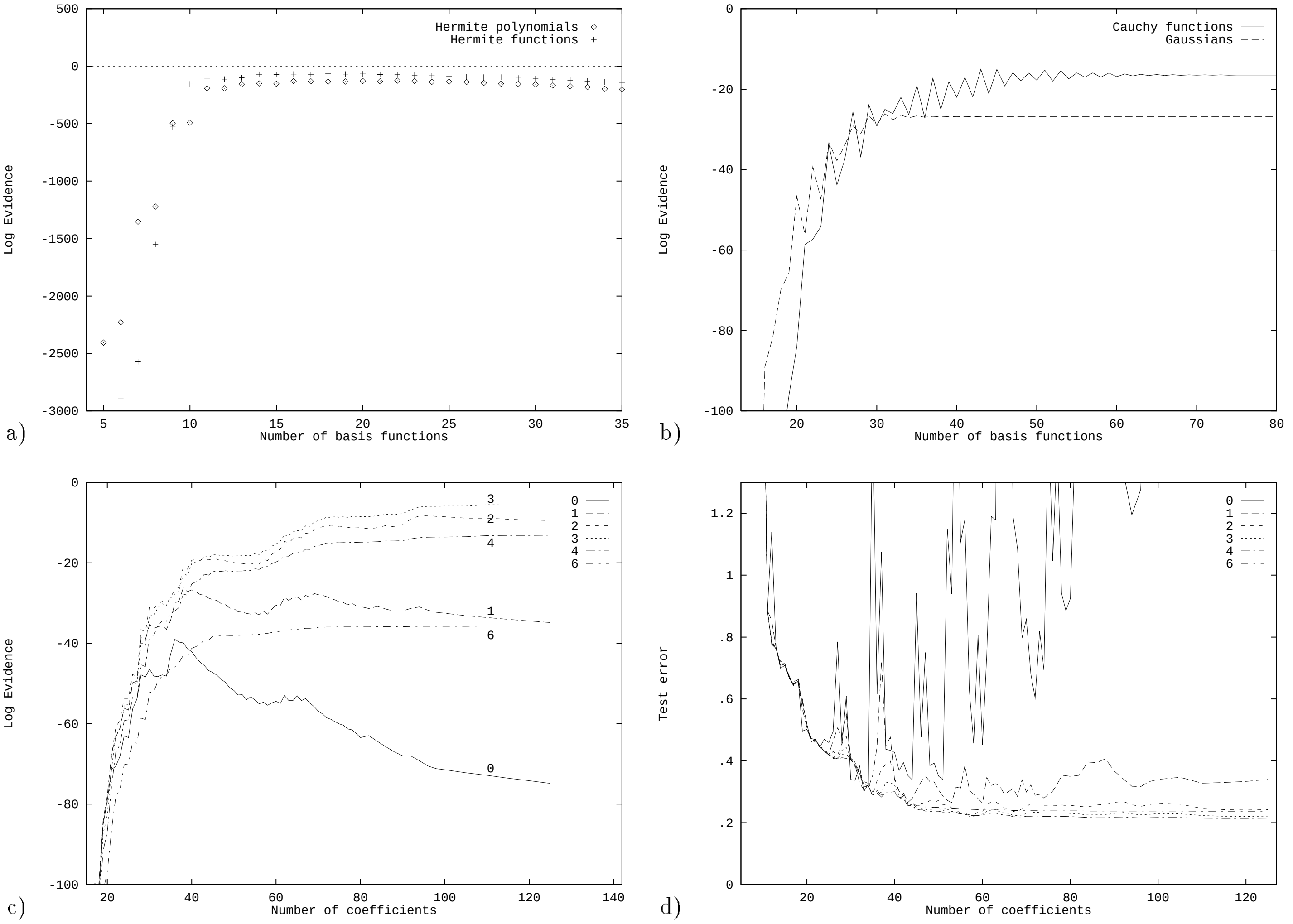


Figure 7: **The Evidence for data set X**

a) Log Evidence for Hermite polynomials and functions. Notice the evidence maximum. The gentle slope to the right is due to the 'Occam factors' which penalise the increasing complexity of the model. **b) Log Evidence for radial basis function models.** Notice that there is no Occam penalty for the additional coefficients in these models, because increased density of radial basis functions does not make the model more powerful. The oscillations in the evidence are due to the details of the pixellation of the basis functions relative to the data points. **c) Log Evidence for splines.** The evidence is shown for the alternative splines regularisers $p = 0 \ldots 6$ (see text). In the representation used, each spline model is obtained in the limit of an infinite number of coefficients. For example, $p = 4$ yields the cubic splines model. **d) Test error for splines.** The number of data points in the test set was 90, *c.f.* number of data points in training set = 37. The y axis shows E_D; the value of E_D for the true interpolant has expectation 0.225 ± 0.02.

6 A Demonstration

These demonstrations will use two one–dimensional data sets, in imitation of [17]. The first data set, 'X,' has discontinuities in derivative (figure 4), and the second is a smoother data set, 'Y' (figure 8). In all the demonstrations, β was not left as a free parameter, but was fixed to its known true value.

The Bayesian method of setting α, assuming a single model is correct, has already been demonstrated, and quantified error bars have been placed on the most probable interpolant (figure 4). The method of evaluating the error bars is to use the posterior covariance matrix of the parameters w_h, $\mathbf{A}^{-1}$, to get the variance on $y(x)$, which for any x is a linear function of the parameters, $y(x) = \sum_h \phi_h(x) w_h$. The error bars at a single point x are given by $\operatorname{var} y(x) = \phi^{\mathrm{T}} \mathbf{A}^{-1} \phi$. However we have access to the full covariance information for the entire interpolant, not just the pointwise error bars. It is possible to visualise the joint error bars on the interpolant by making typical samples from the posterior distribution, performing a random walk around the posterior 'bubble' in parameter space [18]. It is simple to program such a random walk for interpolation problems such as are examined in this paper; however for lack of dynamic paper I will have to leave a demonstration of this to your imagination.

In this section objective comparison of alternative models will be demonstrated; this will be illustrated first with models differing only in the number of free parameters (for example polynomials of different degrees), then with comparisons between models as disparate as splines, radial basis functions and feedforward neural networks. For each individual model, the value of α is optimised, and the evidence is evaluated by integrating over α using the gaussian approximation. All logarithms are to base e.

Hermite polynomials and Hermite functions: Occam's razor for the number of basis functions

Figure 7a shows the evidence for Hermite polynomials of different degrees for data set X. This figure also shows the evidence for Hermite functions, by which I mean Hermite polynomials multiplied by $e^{-x^2/2}$. A regulariser of the form $E_W = \sum \frac{1}{2} 2^{2h} w_h^2$ was used in both cases because the leading polynomial coefficient of Hermite polynomials is 2^h, and it was found that the evidence for the flat regulariser $\sum \frac{1}{2} w_h^2$ was much smaller.

Notice that an evidence maximum is obtained: beyond a certain number of terms, the evidence starts to decrease. This is the Bayesian Occam's razor at work. The additional terms make the model more powerful, able to make more predictions. This power is automatically penalised. Notice the characteristic shape of the 'Occam hill.' On the left, the hill is very steep as the over–simple models fail to fit the data; the penalty for misfitting the data scales as N, the number of data measurements. The other side of the hill is much less steep; the 'Occam factors' here only scale as $k \log N$, where k is the number of parameters. We note in table 1 the values of the maximum evidence achieved by these two models, and move on to alternative models.

Fixed radial basis functions

The basis functions are $\phi_h(x) = g((x - x_h)/r)$, with x_h equally spaced over the range of interest. I will examine two choices of g: a gaussian and a Cauchy function, $1/1 + x^2$. We can quantitatively compare these alternative models of spatial correlation for any data set by evaluating the evidence. The regulariser is $E_W = \sum \frac{1}{2} w_h^2$. Note that this model includes one new free parameter, r; in these demonstrations this parameter has been set to its most probable value (*i.e.* the value which maximises the evidence). To penalise this free parameter an Occam factor is included, $\sqrt{2\pi} \Delta r P(r)$, where $\Delta r =$ posterior uncertainty in r, and $P(r)$ is the prior on r, which is usually subjective to a small degree. This radial basis function model is identical to the 'intrinsic correlation' model of Gull, Skilling and Sibisi [6, 17].

Figure 7b shows the evidence as a function of the number of basis functions, k. Note that for these models there is *not* an Occam penalty for large numbers of parameters. The reason for this is that these extra parameters do not make the model any more powerful (for fixed α and r). The increased density of basis functions does not enable the model to make any significant new predictions because the kernel g band–limits the possible interpolants.

Splines: Occam's razor for the choice of regulariser

I implement the splines model as follows: let the basis functions be a fourier set $\cos hx, \sin hx$, $h = 0, 1, 2, \ldots$. Use the regulariser $E_W = \sum \frac{1}{2} h^p w_{h(\cos)}^2 + \sum \frac{1}{2} h^p w_{h(\sin)}^2$. If $p = 4$ then in the limit $k \to \infty$ we have the cubic splines regulariser $E_y^{(4)} = \int y''(x)^2 dx$; if $p = 2$ we have the regulariser $E_y^{(2)} = \int y'(x)^2 dx$, etc. Figure 7c shows the evidence for data set X as a function of the number of terms, for $p = 0, 1, 2, 3, 4, 6$. Notice that in terms of Occam's razor, both cases discussed above occur: for $p = 0, 1$, as k increases, the model becomes more powerful and there is an Occam penalty. For $p = 3, 4, 6$, increasing k gives rise to no penalty. The case $p = 2$ seems to be on the fence between the two.

As p increases, the regulariser becomes more opposed to strong curvature. Once we reach $p = 6$, the model becomes improbable because the data does in fact have sharp discontinuities. The evidence can choose the order of our splines regulariser for us. For this data set, it turns out that $p = 3$ is the most probable value of p by a few multiples of e.

Results for a smoother data set

Figure 8 shows data set Y, which comes from a much smoother interpolant than data set X. Table 1 summarises the evidence for the alternative models. Note the differences from data set X: in the splines family, the most probable value of p has shifted upwards to the models which penalise curvature more strongly, as we would intuitively expect; among the two radial basis function models, the gaussian correlation model now has a slight edge; Hermite functions, which were a poor model for data set X, are now in first place, for a reason which will become clear shortly.

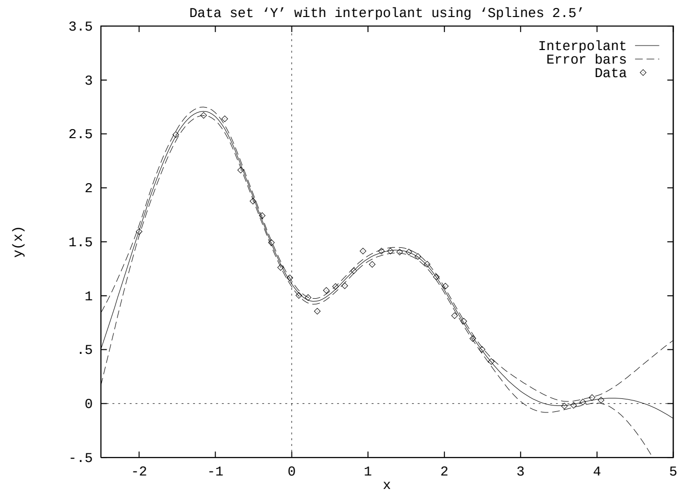


Figure 8: **Data set 'Y', interpolated with splines, $p = 5$.**

Why Bayes can't systematically reject the truth

Let us ask a frequentist question: if one of the hypotheses we offer to Bayes is actually true, *i.e.* it is the model from which the data were generated, then is it possible for Bayes to systematically (over the ensemble of possible data sets) prefer a false hypothesis? Clearly under a worst case analysis, a Bayesian's posterior may favour a false hypothesis. Furthermore, Skilling demonstrated that with some data sets a free form (maximum entropy) hypothesis can have greater evidence than the truth [20], but is it possible for this to happen in the *typical* case, as Skilling seems to claim? I will show that the answer is no, the effect that Skilling demonstrated cannot be systematic. To be precise, the expectation over possible data sets of the log evidence for the true hypothesis is greater than the expectation of the log evidence for any other fixed hypothesis. What then was the cause of Skilling's result? Presumably the particular parameter values of the true model that generated the data were not typical of the prior used when evaluating the evidence for the true model.

Proof. Suppose that the truth is actually $\mathcal{H}_1$. A single data set arrives and we compare the evidences for $\mathcal{H}_1$ and $\mathcal{H}_2$, a different fixed hypothesis. Both hypotheses may have free parameters, but this will be irrelevant to the argument. Intuitively we expect that the evidence for $\mathcal{H}_1$, $P(D|\mathcal{H}_1)$, should usually be greatest. Examine the difference in log evidence between $\mathcal{H}_1$ and $\mathcal{H}_2$. The expectation of this difference, given that $\mathcal{H}_1$ is true, is

$$\left\langle \log \frac{P(D|\mathcal{H}_1)}{P(D|\mathcal{H}_2)} \right\rangle = \int d^N D \, P(D|\mathcal{H}_1) \log \frac{P(D|\mathcal{H}_1)}{P(D|\mathcal{H}_2)}.$$

(Note that this integral implicitly integrates over all $\mathcal{H}_1$'s parameters according to their prior distribution under $\mathcal{H}_1$.) Now it is well known that for normalised p and q, $\int p \log \frac{p}{q}$ has a minimum value over q of 0, which is only achieved by setting $q = p$. Therefore a distinct hypothesis $\mathcal{H}_2$ is never expected to *systematically* defeat the true hypothesis, for just the same reason that it is not wise to bet differently from the true odds. ●

This has two important implications. First, it gives us frequentist confidence in the ability of Bayesian methods on the average to identify the true hypothesis. Secondly, it provides a severe test of any numerical implementation of a Bayesian inference framework: imagine that we have written a program that evaluates the evidence for hypotheses $\mathcal{H}_1$ and $\mathcal{H}_2$; then we can generate mock data from sources simulating $\mathcal{H}_1$ and $\mathcal{H}_2$ and evaluate the evidences; if there is any systematic bias, averaged over several mock data sets, for the estimated evidence to favour the false hypothesis, then we can be sure that our numerical implementation is not evaluating the evidence correctly.

This issue is illustrated using data set Y. The 'truth' is that this data set was actually generated from a quadratic Hermite function, $1.1(1-x+2x^2)e^{-x^2/2}$. By the above argument the evidence ought probably to favour the hypothesis 'the interpolant is a 3–coefficient Hermite function' over our other hypotheses. Let us evaluate the evidence for a variety of hypotheses and confirm that none of them has greater evidence than the true hypothesis. Table 1 shows the evidence for the true Hermite function model, and for other models. Notice that the truth is indeed considerably more probable than the alternatives.

Having demonstrated that Bayes cannot systematically fail when one of the hypotheses is true, we now examine the way in which this framework can fail, if none of the hypotheses offered to Bayes is any good.

Comparison with 'generalisation error'

It is a popular and intuitive criterion for choosing between alternative interpolants to compare their errors on a test set that was not used to derive the interpolant. 'Cross–validation' is a more refined version of this same idea. How does this method relate to the evaluation of the evidence described in this paper?

Figure 7c displayed the evidence for the family of spline interpolants. Figure 7d shows the corresponding test error, measured on a test set with size over twice as big (90) as the 'training' data set (37) used to determine the interpolant. A similar comparison was made in figure 5b. Note that the overall trends shown by the evidence are matched by trends in the test error (if you flip one graph upside down). Also, for this particular problem, the ranks of the alternative spline models under the evidence are similar to their ranks under the test error. And in figure 5b, the evidence maximum was surrounded by the test error minima. Thus this suggests that the evidence might be a reliable predictor of generalisation ability. However, this is not necessarily the case. There are five reasons why the evidence and the test error might not be correlated.

First, the test error is a noisy quantity. It is necessary to devote large quantities of data to the test set to obtain a reasonable signal to noise ratio. In figure 5b more than twice as much data is in each test set but the difference in

Table 1: **Evidence for data sets X and Y**

Model	Data Set X		Data Set Y	
	Best parameter values	Log evidence	Best parameter values	Log evidence
Hermite polynomials	$k = 22$	-126	$k = 9$	1.1
Hermite functions	$k = 18$	-66	$k = 3$	42.2
Gaussian radial basis functions	$k > 40$, $r = .25$	-28.8 ± 1.0	$k > 50$, $r = .77$	27.1 ± 1.0
Cauchy radial basis functions	$k > 50$, $r = .27$	-18.9 ± 1.0	$k > 50$, $r = 1.1$	25.7 ± 1.0
Splines, $p = 2$	$k > 80$	-9.5	$k > 50$	8.2
Splines, $p = 3$	$k > 80$	-5.6	$k > 50$	19.8
Splines, $p = 4$	$k > 80$	-13.2	$k > 50$	22.1
Splines, $p = 5$	$k > 80$		$k > 50$	21.8
Splines, $p = 6$	$k > 80$	-35.8	$k > 50$	20.4
Neural networks	8 neurons, $k = 25$	-12.6	6 neurons, $k = 19$	25.7

α between the two test error minima exceeds the size of the Bayesian confidence interval for α.

Second, the model with greatest evidence is not expected to be the best model all the time — Bayesian inferences are uncertain. The whole point of Bayes is that it quantifies precisely those uncertainties: the relative values of the evidence for alternative models express the plausibility of the models, given the data and the underlying assumptions.

Third, there is more to the evidence than there is to the generalisation error. For example, imagine that for two models, the most probable interpolants happen to be identical. In this case, the generalisation error for the two solutions must be the same. But the evidence will not in general be the same: typically, the model that was *a priori* more complex will suffer a larger Occam factor and will have a smaller evidence.

Fourth, the test error is a measure of performance only of the single most probable interpolant: the evidence is a measure of plausibility of the entire posterior ensemble around the best fit interpolant. A stronger correlation between the evidence and the test statistic would be obtained if the test statistic used were the average of the test error over the posterior ensemble of solutions. This ensemble test error is not so easy to compute.

The fifth and most interesting reason why the evidence might not be correlated with the generalisation error is that there might be a flaw in the underlying assumptions such that the hypotheses being compared might all be poor hypotheses. If a poor regulariser is used, for example, one that is ill–matched to the statistics of the world, then the Bayesian choice of α will often not be the best in terms of generalisation error [3, 6, 9]. Such a failure occurs in the companion paper on neural networks. What is our attitude to such a failure of Bayesian prediction? The failure of the evidence does not mean that we should discard the evidence and use the generalisation error as our criterion for choosing α. A failure is an opportunity to learn; a healthy scientist searches for such failures, because they yield insights into the defects of the current model. The detection of such a failure (by evaluating the generalisation error for example) motivates the search for new hypotheses which do not fail in this way; for example alternative regularisers can be tried until a hypothesis is found that makes the data more probable.

If one only uses the generalisation error as a criterion for model comparison, one is denied this mechanism for learning. The development of image deconvolution was held up for decades because no–one used the Bayesian choice of α; once the Bayesian choice of α was used [6], the results obtained were most dissatisfactory, making clear what a poor regulariser was being used; this motivated an immediate search for alternative priors; the new priors discovered by this search are now at the heart of the state of the art in image deconvolution.

Admitting neural networks into the canon of Bayesian interpolation models

A second paper will discuss how to apply this Bayesian framework to the task of evaluating the evidence for feedforward neural networks. Preliminary results using these methods are included in table 1. Assuming that the approximations used were valid, it is interesting that the evidence for neural nets is actually good for both the spiky and the smooth data sets. Furthermore, neural nets, in spite of their arbitrariness, yield a relatively compact model, with fewer parameters needed than to specify the splines and radial basis function solutions.

7 Conclusions

The recently developed methods of Bayesian model comparison and regularisation have been presented. Models are ranked by evaluating the evidence, a solely data–dependent measure which intuitively and consistently combines a model's ability to fit the data with its complexity.

Regularising constants are set by maximising the evidence. For many regularisation problems, the theory of the number of well determined parameters makes it possible to perform this optimisation on–line.

In the interpolation examples discussed, the evidence was used to set the number of basis functions k in a polynomial model; to set the characteristic size r in a radial basis function model; to choose the order p of the regulariser for a spline model; and to rank all these different models in the light of the data.

Further work is needed to formalise the relationship of this framework to the pragmatic model comparison technique of cross–validation. By using the two techniques in parallel it is possible to detect flaws in the underlying assumptions implicit in the data models being used. Such failures direct us in our search for superior models, providing a powerful tool for human learning. There are thousands of data modelling tasks waiting for the evidence to be evaluated for them. It will be exciting to see how much we can learn when this is done.

References

[1] J. Berger (1985). *Statistical decision theory and Bayesian analysis,* Springer.

[2] G.L. Bretthorst (1990). Bayesian Analysis. I. Parameter Estimation Using Quadrature NMR Models. II. Signal Detection and Model Selection. III. Applications to NMR., *J. Magnetic Resonance* **88** 3, 533–595.

[3] A.R. Davies and R.S. Anderssen (1986). Optimization in the regularization of ill–posed problems, *J. Austral. Mat. Soc. Ser. B* **28**, 114–133.

[4] W.T. Grandy, Jr., editor (1991). *Maximum Entropy and Bayesian Methods, Laramie 1990,* Kluwer.

[5] S.F. Gull (1988). Bayesian inductive inference and maximum entropy, in *Maximum Entropy and Bayesian Methods in science and engineering, vol. 1: Foundations*, G.J. Erickson and C.R. Smith, eds., Kluwer.

[6] S.F. Gull (1989). Developments in Maximum entropy data analysis, in [19], 53–71.

[7] S.F. Gull (1989). Bayesian data analysis: straight–line fitting, in [19], 511–518.

[8] S.F. Gull and J. Skilling (1991). *Quantified Maximum Entropy.* `MemSys5` *User's manual,* M.E.D.C., 33 North End, Royston, SG8 6NR, England.

[9] D. Haussler, M. Kearns and R. Schapire (1991). Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension, Preprint.

[10] E.T. Jaynes (1986). Bayesian methods: general background, in *Maximum Entropy and Bayesian Methods in Applied Statistics,* ed. J.H. Justice, C.U.P.

[11] H. Jeffreys (1939). *Theory of Probability,* Oxford Univ. Press.

[12] T.J. Loredo (1989). From Laplace to supernova SN 1987A: Bayesian inference in astrophysics, in *Maximum Entropy and Bayesian Methods,* ed. P. Fougere, Kluwer.

[13] D.J.C. MacKay (1991). A practical Bayesian framework for backprop networks, submitted to *Neural computation.*

[14] R.M. Neal (1991). Bayesian mixture modeling by Monte Carlo simulation, Preprint.

[15] T. Poggio, V. Torre and C. Koch (1985). Computational vision and regularization theory, *Nature* **317** 6035, 314–319.

[16] G. Schwarz (1978). Estimating the dimension of a model, *Ann. Stat.* **6** 2, 461–464.

[17] S. Sibisi (1990). Bayesian interpolation, in [4], 349–355.

[18] J. Skilling, D.R.T. Robinson, and S.F. Gull (1991). Probabilistic displays, in [4], 365–368.

[19] J. Skilling, editor (1989). *Maximum Entropy and Bayesian Methods, Cambridge 1988,* Kluwer.

[20] J. Skilling (1991). On parameter estimation and quantified MaxEnt, in [4], 267–273.

[21] J. Skilling (1989). The eigenvalues of mega–dimensional matrices, in [19], 455–466.

[22] R. Szeliski (1989). *Bayesian modeling of uncertainty in low level vision,* Kluwer.

[23] D. Titterington (1985). Common structure of smoothing techniques in statistics, *Int. Statist. Rev.* **53**, 141–170.

[24] C. S. Wallace and D. M. Boulton (1968). An information measure for classification, *Comput. J.* **11** 2, 185–194.

[25] C. S. Wallace and P. R. Freeman (1987). Estimation and Inference by Compact Coding, *J. R. Statist. Soc. B* **49**3, 240-265.

I thank Mike Lewicki, Nick Weir and David R.T. Robinson for helpful conversations, and Andreas Herz for comments on the manuscript. I am grateful to Dr. R. Goodman and Dr. P. Smyth for funding my trip to Maxent 90.

This work was supported by a Caltech Fellowship and a Studentship from SERC, UK.