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. It will be exciting to see how much we can learn when this is done.

was held up for years because no–one used the Bayesian choice of α; once the Bayesian choice of α was used [22], 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, more probable priors discovered by this search are now at the heart of the state of the art in image deconvolution [76].

The similarity between regularisation and 'early stopping'

While an over–parameterised model is fitted to a data set using gradient descent on the data error, it is sometimes noted that the model's generalisation error passes through a minimum, rather than decreasing monotonically. This is known as 'over–learning' in the neural networks community, and some researchers advocate the use of 'early stopping,' that is, stopping gradient descent before the data error minimum is reached, so as to try to obtain solutions with smaller generalisation error.

This author believes that 'over–learning' should be viewed as a symptom of a model ill–matched to the data set, and that the appropriate response is not to patch up a bad model, but rather to search for models which are well matched to our data. In particular, the use of models incorporating simple regularisers is expected to give results qualitatively similar to early stopping. This can be seen by examining figure 2.6. The regulariser moves the minimum of the objective function from $\mathbf{w}_{\rm ML}$ to $\mathbf{w}_{\rm MP}$; as the strength of the regulariser α is increased, $\mathbf{w}_{\rm MP}$ follows a knee–shaped trajectory from $\mathbf{w}_{\rm ML}$ to the origin; a typical solution $\mathbf{w}_{\rm MP}$ is shown in figure 2.6. If on the other hand gradient descent on the likelihood (data error) is used, and if the typical initial condition is close to the origin, then gradient descent will follow a similar knee–shaped trajectory. Thus qualitatively similar solutions are expected from increasingly early stopping and increasingly strong regularisation with complete minimisation. Regularisation is to be preferred as a more robust, repeatable and comprehensible procedure.

Admitting neural networks into the canon of Bayesian interpolation models

Chapter 3 will discuss how to apply this Bayesian framework to feedforward neural networks. Preliminary results using these methods are included in table 2.1. 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.

2.7 Conclusions

The recently developed methods of Bayesian model comparison and regularisation have been presented. Models can be 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. The precise posterior probabilities of the models also depend on the subjective priors that we assign to them, but these terms are typically overwhelmed by the evidence.

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

Comparison with 'generalisation error'

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

Figure 2.7c displayed the evidence for the family of spline interpolants. Figure 2.7d 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 2.5b. 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 2.5b, the evidence maximum over α 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 2.5b more than twice as much data is in each test set but the difference in $\log \alpha$ between the two test error minima exceeds the size of the Bayesian confidence interval for $\log \alpha$.

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 two solutions will have the same generalisation error, 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. Probably 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 models being compared might all be poor models. 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 [13, 22, 26]. Such a failure occurs in chapter 3. What is our attitude to such a failure of Bayesian prediction? The failure of the evidence does not mean that we should discard Bayes' rule and use the generalisation error as our criterion for choosing α. A failure is an opportunity to learn; a healthy scientist actively 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 models which do not fail in this way; for example alternative regularisers can be tried until a model 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 maximum entropy image deconvolution

Why Bayes can't systematically reject the truth

Let us ask a sampling theory question: if one of the models 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 model? Clearly under a worst case analysis, a Bayesian's posterior may favour a false model. Furthermore, Skilling demonstrated that with some data sets a free form (maximum entropy) model can have greater evidence than the truth [63]; 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, that the effect that Skilling demonstrated cannot be systematic. To be precise, the expectation over possible data sets of the log evidence for the true model is greater than the expectation of the log evidence for any other fixed model [49].[14]

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 model. Both models 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. Let us 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, $\int p \log \frac{p}{q}$ is minimised by setting $q = p$ (Gibb's theorem). Therefore a distinct model $\mathcal{H}_2$ is never expected to *systematically* defeat the true model, 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 confidence in the ability of Bayesian methods on the average to identify the true model. Secondly, it provides a stringent test of numerical implementations of Bayesian model comparison. Imagine that we have written a program that evaluates the evidence for models $\mathcal{H}_1$ 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 model, 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 model 'the interpolant is a 3–coefficient Hermite function' over our other models. Table 2.1 shows the evidence for the true Hermite function model, and for other models. As already stated, the truth is indeed considerably more probable than the alternatives.

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

[14] Skilling's result presumably occurred because the particular parameter values of the true model that generated the data were not typical of the prior used when evaluating the evidence for that model. In such a case, the log evidence difference can show a transient bias against the true model, for small quantities of data; such biases are usually corrected by greater quantities of data.

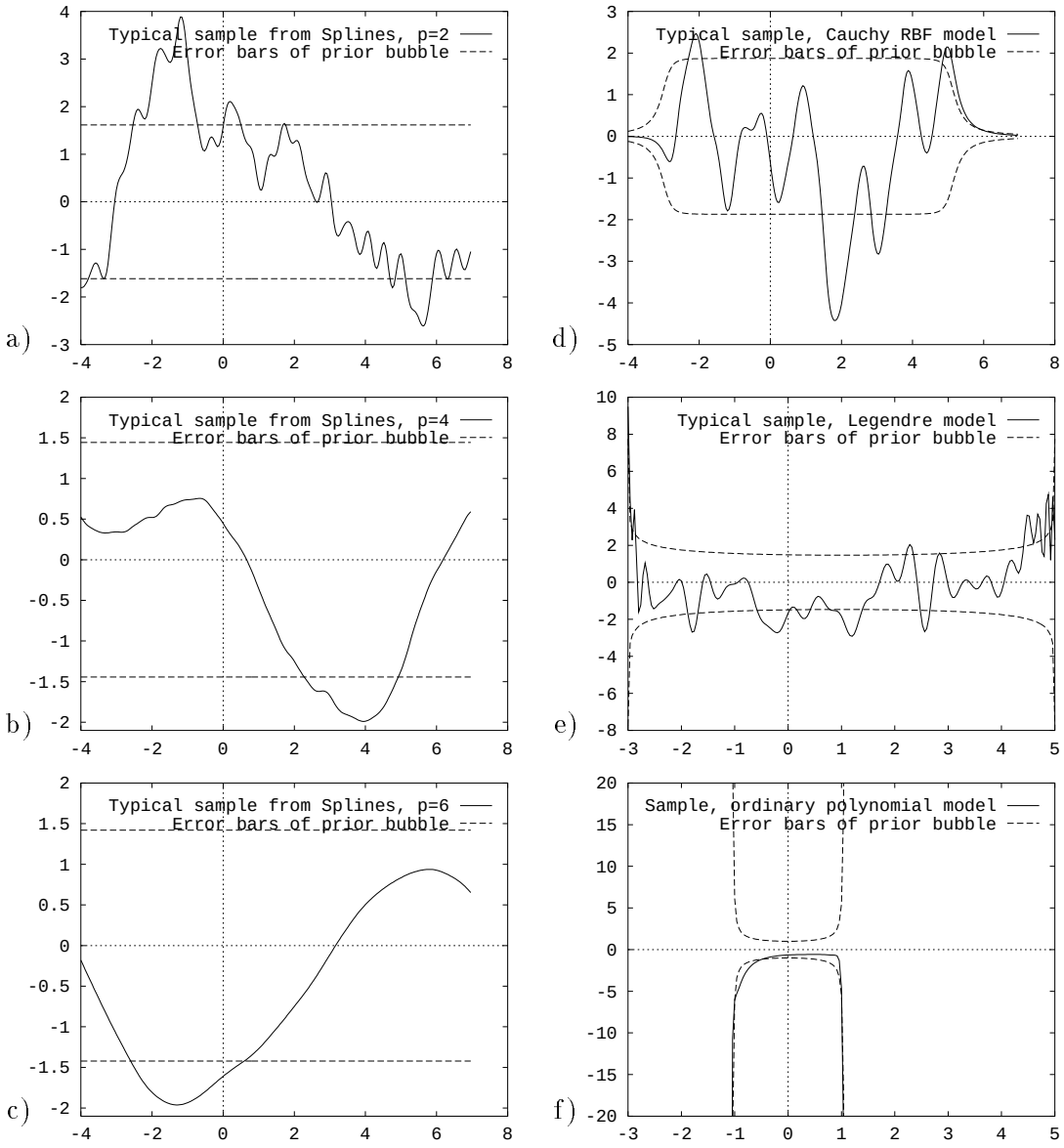

Figure 2.9: **Typical samples from the prior distributions of six models**
This figure illustrates the character of some of the alternative models used in this chapter. Each model was represented with 60 basis functions, and a typical sample from the prior distribution is shown. The regularisation constant was set in each case so that the typical magnitude of the interpolant was similar. a) Splines, $p = 2$. b) Splines, $p = 4$ (cubic splines). c) Splines, $p = 6$. The splines were represented with a Fourier set with period 12.0. Notice how the spikiness of the typical sample decreases as the order of the spline model increases. d) Cauchy radial basis functions. The basis functions were equally spaced from -3.0 to 5.0, and had scale $r = 0.2975$. e) Legendre polynomials. The polynomials were stretched so that the interval [-3.0,5.0] corresponds to the natural interval. Notice that the characteristic amplitude diverges at the boundaries, and the characteristic frequency of the typical sample also increases towards the boundaries. f) Ordinary polynomials. This figure illustrates what hideous results can be obtained if a prior is carelessly assigned. A uniform prior over the coefficients of $y = \sum w_h x^h$ yields a highly non–uniform typical sample.

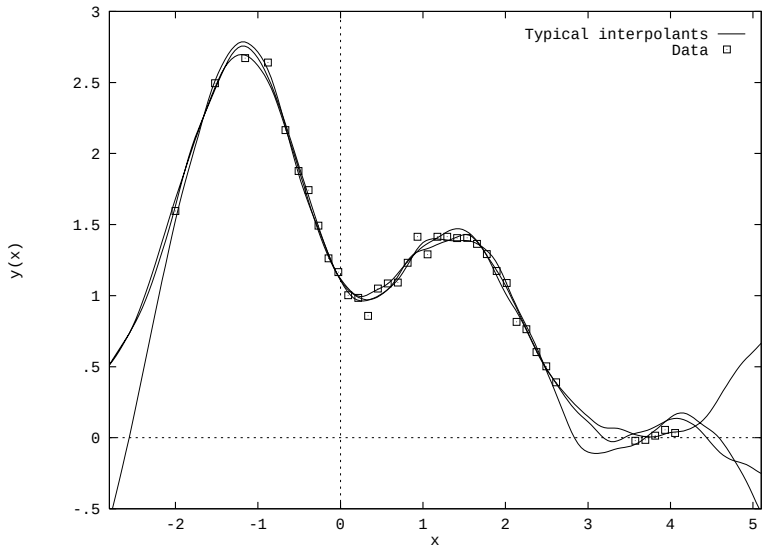

Figure 2.8: **Data set 'Y,' interpolated with splines,** $p = 5$**.**
The data set is shown with three typical interpolants drawn from the posterior probability distribution. Contrast this with figure 2.4b, in which the most probable interpolant is shown with its pointwise error bars.

out that $p=3$ is the most probable value of p, by a few multiples of e.

In passing, the radial basis function models described above can be transformed into the splines models' Fourier representation. If the radial basis function kernel is $g(x)$ then the regulariser in the splines representation is $E_W = \sum \frac{1}{2}(w^2_{h(\cos)}+w^2_{h(\sin)})G_h^{-2}$, where G_h is the discrete Fourier transform of g.

Results for a smoother data set

Figure 2.8 shows data set Y, which comes from a much smoother interpolant than data set X. Table 2.1 summarises the evidence for the alternative models. We can confirm that the evidence behaves in a reasonable manner by noting the following differences between data sets X and Y:

In the splines family, the most probable value of p has shifted upwards to the stiffer splines with $p=4-5$, as we would intuitively expect.

Legendre polynomials: an observant reader may have noticed that when data set X was modelled with Legendre polynomials, the most probable number of coefficients $k=38$ was suspiciously similar to the number of data points $N=37$. For data set Y, however, the most probable number of coefficients is 11, which confirms that the evidence does not always prefer the polynomial with $k=N$. Data set X behaved in this way because it is very poorly modelled by polynomials.

The Hermite function model, which was a poor model for data set X, is now the most probable, by a long way (over a million times more probable). The reason for this is that actually the data *were* generated from a Hermite function!

Table 2.1: **Evidence for models interpolating data sets X and Y**
All logs are natural. The evidence $P(D|\mathcal{H})$ is a density over D space, so the absolute value of the log evidence is arbitrary within an additive constant. Only differences in values of log evidences are relevant, relating directly to probability ratios.

Model	Data Set X		Data Set Y	
	Best parameter values	Log evidence	Best parameter values	Log evidence
Legendre polynomials	$k = 38$	-47	$k = 11$	23.8
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, $p = 5$	$k > 80$	-24.9	$k > 50$	21.8
Splines, $p = 6$	$k > 80$	-35.8	$k > 50$	20.4
Hermite functions	$k = 18$	-66	$k = 3$	42.2
Neural networks	8 neurons, $k = 25$	-12.6	6 neurons, $k = 19$	25.7

Sibisi [11, 22, 60].

Figure 2.7b shows the evidence as a function of the number of basis functions, k. Note that for these models there is *not* an increasing 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

The splines models were implemented 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. Notice that the 'non–parametric' splines model can easily be put in an explicit parameterised representation. Note that none of these splines models include 'knots.'

Figure 2.7c 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 less probable because the data demand sharp discontinuities. The evidence can choose the order of our splines regulariser for us. For this data set, it turns

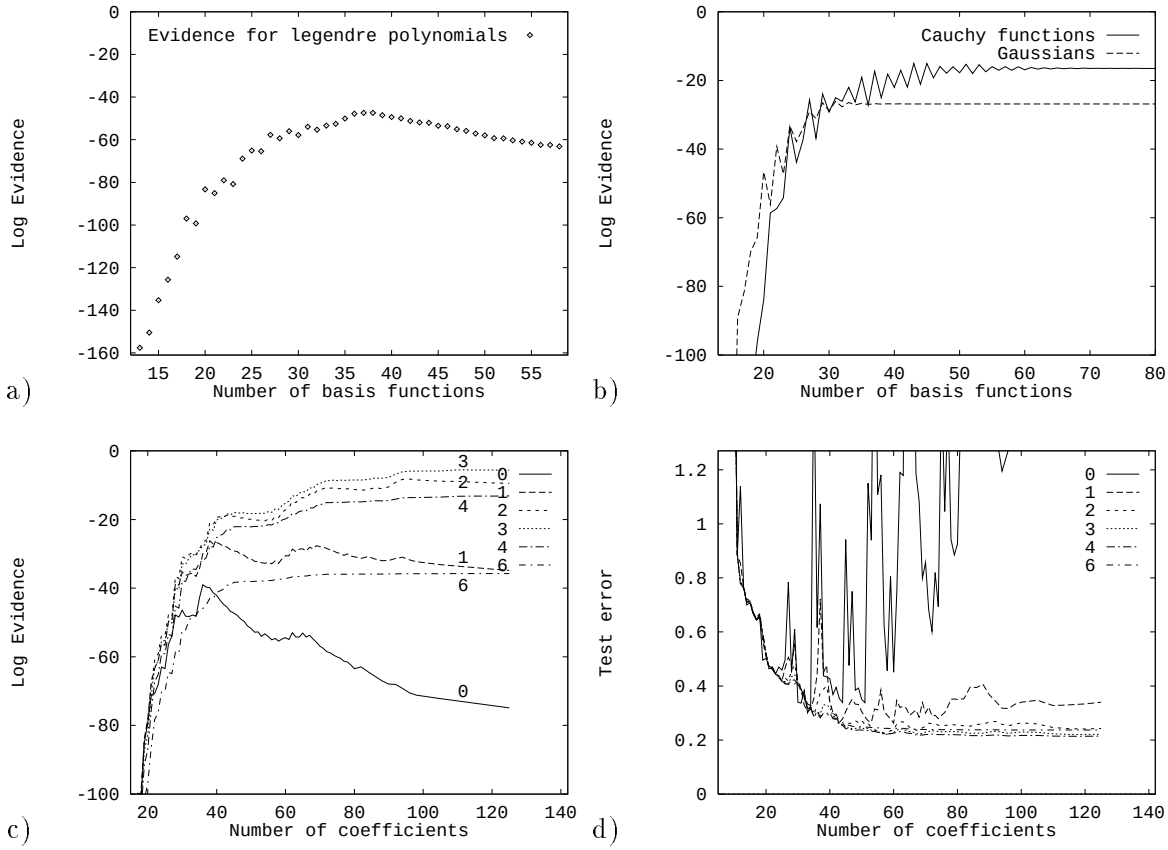

Figure 2.7: **The Evidence for data set X** (see also table 1)
a) Log Evidence for Legendre polynomials. Notice the evidence maximum. The gentle slope to the right is due to the 'Occam factors' which penalise the increasing complexity of the model. **b) Log Evidence for radial basis function models.** 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.