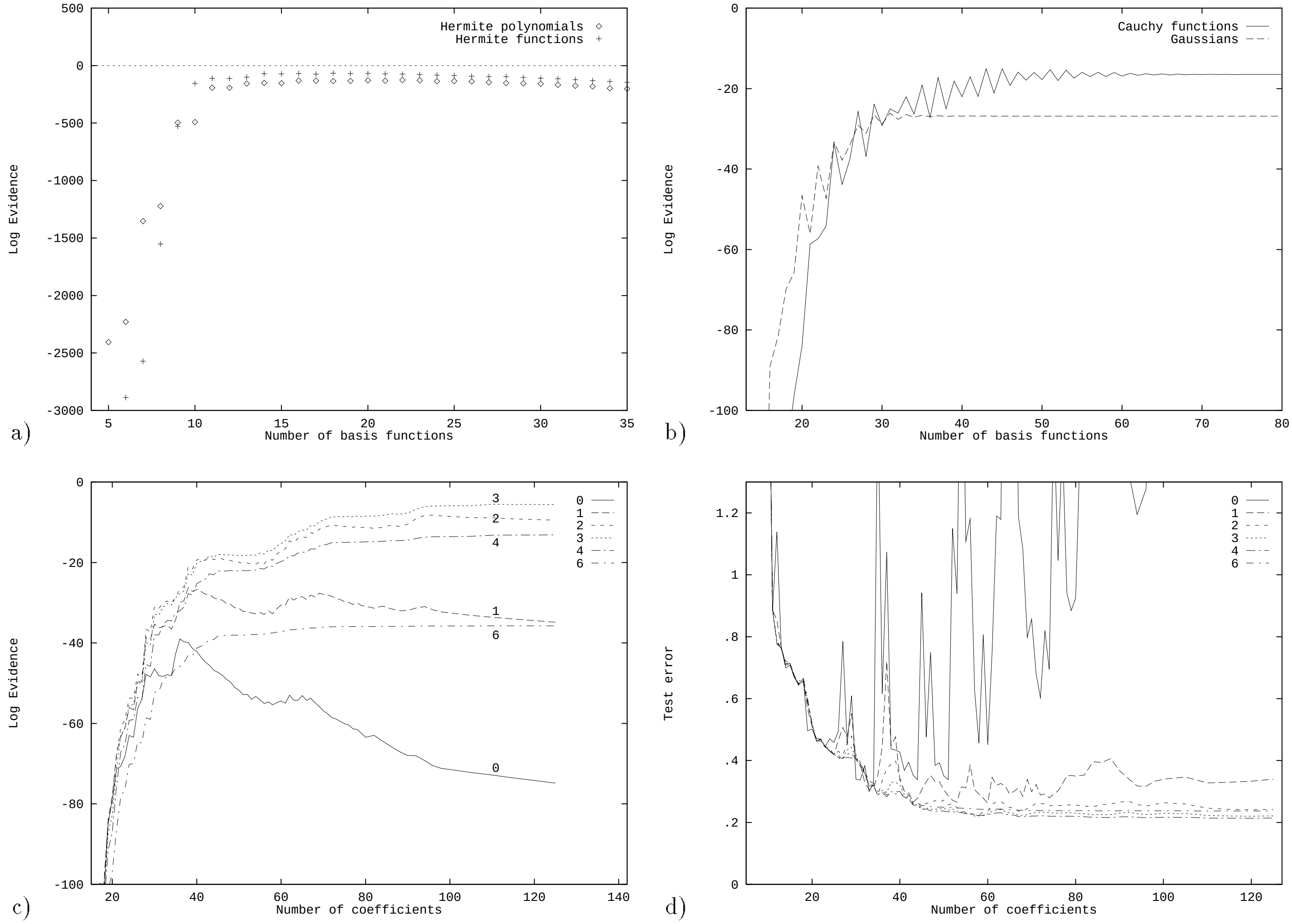

Figure 7: **The Evidence for data set X**

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.

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 Δr = 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.

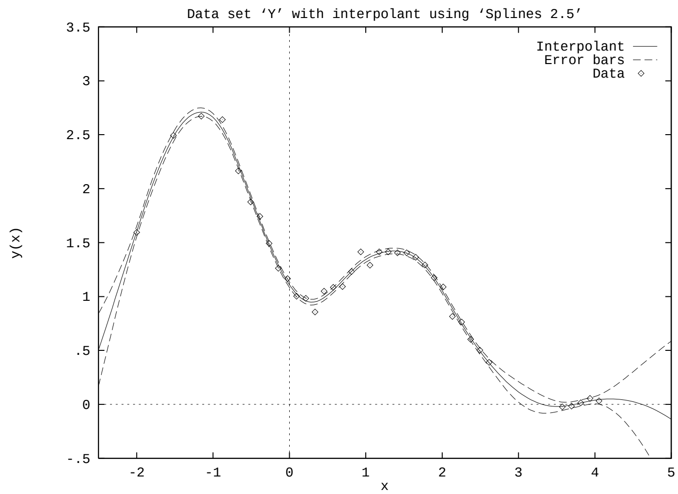

Figure 8: **Data set 'Y', interpolated with splines, $p = 5$.**

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 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.

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 $\pm$ 1.0	$k > 50$, $r = .77$	27.1 $\pm$ 1.0
Cauchy radial basis functions	$k > 50$, $r = .27$	-18.9 $\pm$ 1.0	$k > 50$, $r = 1.1$	25.7 $\pm$ 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

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* **493**, 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.