Chapter 7

Postscript

It is common that, following several years' devotion to a religion, a student's views about that religion will have matured. This postscript is intended to communicate the reservations and criticisms I now have about the Bayesian methods described, and the open questions and problems that remain.

7.1 The closed hypothesis space

Bayesian hard–liners often thump the 'Cox axioms' drum, proclaiming that consistent inference can only be Bayesian; but it is rarely made clear what Cox's axioms are. In fact Cox's result assumes (among other things) that we are performing inductive inference in a defined closed hypothesis space.

This is a two edged sword. The good aspect is that Bayesian inductive inference cannot proceed until all properties of our hypothesis space have been articulated; thus we are forced to make explicit all our assumptions. Furthermore, once the hypothesis space is defined, Bayesian inference is a mechanical and well–defined process. Bayesianism does not need to consult sampling theory criteria such as 'efficiency,' 'unbiasedness,' 'consistency,' 'sufficiency,' 'uniform convergence,' or 'minimum variance' — desiderata which can often be mutually conflicting! We simply write down the conditional assumptions that we are making (for example the data that have occurred), and the propositions whose plausibilities we wish to infer, and evaluate $P(\text{Proposition}|\text{Assumptions})$ by using the sum and product rules of probability within the defined hypothesis space. The axioms on which probabilistic inference are based guarantee that inferences made in this way will be coherent.

On the other hand, several deficiencies arise from the constraint of a closed hypothesis space. The central problem is that our Bayesian inferences are obtained assuming that the hypothesis space is right, but we have no Bayesian way of assessing whether our hypothesis space *is* right, apart from coming up with alternative hypothesis spaces with which comparisons can be made.

For example, the error bars discussed throughout this thesis are evaluated assuming that the model is true. I do not think that a non–Bayesian procedure can improve on this (orthodox confidence intervals are identical to Bayesian error bars in the gaussian case), but it is important to be aware that it is possible for the true interpolant to lie well outside the error bars assigned by a model, if that model is defective in some way. An example of this can be seen in figure 2.4b, where the error bars fail to include a point that in fact was not an outlier. The Bayesian resolution of this is to examine other models; when the radial basis function model in figure 2.4 is replaced by a more probable neural network model

(Table 2.1), the interpolant goes much closer to this data point (which is probably why the neural network model is more probable).

A second defect of the closed hypothesis space assumption is discussed in Chapter 4. The expected information gain provided by a datum was defined, like the error bars, by assuming that the model is correct. In extreme cases this may lead to ludicrous results — distant data points may be evaluated as mutually informative because of a misleading interdependence in the model. In Chapter 5 it is shown that the mean marginal expected information gain does seem to be marred by artifacts arising from the assumption that the model is correct.

Thirdly, Bayes' rule provides no mechanism for 'alternative–free tests' of a hypothesis space. This reservation about Bayesian methods has also been expressed by Lindley [44]. Hard line Bayesians would retort that there is no such thing as an alternative–free test, and certainly most of the classical alternative–free tests do have an implicit alternative. For example, a classical test of a parameter being zero has as an implicit alternative the hypothesis that the parameter has a value in an interval with a derivable prior width [38]. But I do believe that we perform alternative–free tests. Often, we become dissatisfied with a theory because it seems to be making unusually poor predictions. This prompts us to start searching for superior theories. Without any alternative being more than very vaguely specified, we are able to *infer* that something is wrong. Once we find a superior alternative, we can *then* come back and use Bayes' rule to reject the original theory; but the initial decision to search for a new theory was alternative–free and could not be made with Bayes' rule alone. Nor can Bayes' rule alone direct you towards new models. The invention of hypothesis spaces remains a human's domain.

Having recognised that Bayes' rule cannot perform the alternative–free inferences that are part of the modelling process (the right–hand loop of figure 1.1), I would end on a more optimistic note. I believe that Bayesian methods, *together* with traditional methods such as cross–validation, yield a powerful tool for alternative–free hypothesis testing and new model formation. This was illustrated in Chapter 3, where the poor correlation between the evidence and the test error highlighted an inconsistency in the model space; if one only used the test error (cross–validation) for model comparison, this opportunity for learning would be lost; likewise, if only the Bayesian evidence were evaluated, we would be none the wiser. I think that Bayesians would do well to include similar alternative–free 'warning bells' in their algorithms.

7.2 For approximation, are probabilities relevant?

The Bayesian approach to model comparison evaluates how *probable* alternative models are given the data. In fields such as image reconstruction and NMR, this may be precisely the right thing to do. In contrast, in adaptive modelling, the real problem is often to estimate *how well each model is expected to generalise.* We know perfectly well that the truth is *not* that the data were generated by some neural network whose parameters we now wish to infer! We know that all the models are false, so the Bayesian assessment of the relative probabilities of alternative parameterised models seems almost irrelevant to what we are interested in, which is how well each of the models approximates. Really the Bayesian solution to this task would be to use a model that we really believe in to *infer* what the truth might be, then use *decision theory* to select from the false models the one that is expected to minimse the appropriate cost function.

The startling fact is that in spite of this, the evidence for the false models does seem to be well correlated with generalisation ability, when the model space is well–matched to the problem (figures 3.12 and 5.5). There are theories which attempt to directly predict generalisation ability, leading to Akaike's FPE criterion, and Moody's GPE [3, 45].[1] But for the toy problems I have studied, neither of these criteria has a better correlation with the generalisation error than that achieved by the evidence. It is an open question whether these results are flukes. More work is definitely called for on the relationship between the evidence, cross–validation and generalisation ability to understand these results.

7.3 Having to make too much explicit

Statistical problems that are precisely enough stated for the sampling theory school can be too vague for a Bayesian [44]. When the Bayesian adds additional constraints to such a problem to make it solveable, he comes under fire for making assumptions that may be, in detail, not justified. 'Order statistics' provide an example of such a problem.

Imagine that we wish to infer the median of a density given N samples from it, without a precise specification of what type of density we are dealing with (in particular, we do not know the distribution is Gaussian). A Bayesian analysis would have to assume an explicit parameterised form for the density (for example a free form density with a maxent prior), solve for the posterior distribution of the parameters, then marginalise over that distribution to get a posterior for the median of the density. I think that in principle this is the right thing to do (and, to their credit, some Bayesians have shown that it can be done [11]), but it is an approach that involves introducing and then eliminating a large amount of irrelevant baggage (the explicit parameterised form for the density). The details of this parameterisation will probably be hard to justify, and anyway the answer that we obtain is unlikely to depend sensitively on them. It seems unfortunate to have to introduce so much explicit and arbitrary detail in order to answer a simple question. Having said this, I should make it clear that I am not advocating the orthodox sampling theory approach to order statistics; like most sampling theory procedures, order statistics are incoherent. It will be interesting if Bayesian methods can be developed which avoid having to explicitly handle detailed parameterisations that are then marginalised away again. Perhaps Monte Carlo methods like Radford Neal's [46] are a step in this direction.

7.4 An alternative interpretation of weight decay

Geoff Hinton (personal communication) has suggested an alternative view of mixture weight decay. The decay mechanism is still viewed as implementing prior knowledge, but not the literal prior that says the w_i are modelled as coming from a mixture of gaussians. Rather, our real prior assumption is that a fraction of the weights ought to be exactly zero. Thus the true width of the component at the origin ought to be zero; it is only set to a non–zero value as a computational artifice. This view, that weight decay is intended to switch off weights, is apparently shared by other workers [32, 77].

Under this interpretation, there is no reason to suppose that the Bayesian choice of the width of this component should be appropriate.[2] (The width of the other broad component(s) of the mixture distribution should still be inferred using Bayesian methods.) It will

[1] One aspect of these criteria that I wish to examine is whether, as their authors claim, they predict generalisation ability *without* making implicit assumptions about the properties of the 'truth.'

[2] All the same, Nowlan and Hinton have applied the Bayesian procedure to networks predicting sunspot time series, and obtained better performance than any published model [48].

be interesting to see if this interpretation can be formalised, leading to an alternative objective procedure for setting the parameters of a zero mixture. This would also necessitate changes in the evaluation of the evidence.

7.5 Future tasks, open problems

More expensive evidence calculations, cheaper Hessian calculations

The Bayesian calculations throughout this thesis all depend on the inverse Hessian $\mathbf{A}^{-1}$, under the gaussian approximation. There are two directions for further work. In the more expensive direction, we can ask how to make the Bayesian calculations more accurate by improving on the gaussian approximation, by the use of Monte Carlo methods, for example. In addition, methods for integrating over regularisation constants need to be developed, rather than fixing those constants to their most probable values. These are questions that Skilling and Sibisi (personal communication) are working on.

In the cheaper direction, we can ask how to make the Hessian calculations more approximate and more efficient so as to reduce the cost of these calculations. I hope to investigate statistical methods for reducing the $O(Nk^2)+O(k^3)$ calculation of properties of $\mathbf{A}^{-1}$ to $O(k^2)$ or less time.

Noisy input variables

The tasks of interpolation and classification given noisy input variables has yet to be integrated into the evidence framework.

Missing data

Imagine that we are asked to interpolate a data set $D = \{\mathbf{x}, \mathbf{t}\}$, in which some of the elements $\mathbf{x}$ are incomplete, lacking a specification of some of their components. The interpolation framework described in this thesis cannot handle this case because the density over $\mathbf{x}$ is not modelled. It is an open problem to find a simple, well-defined way to integrate data with missing components into these interpolation models.

This problem has much in common with the task of combining unlabelled data with discriminative training. In discriminative training we adapt a classifier that models $P(\mathbf{t}|\mathbf{x}, \mathbf{w}, \mathcal{H})$ to a data set $D_1 = \{\mathbf{x}, \mathbf{t}\}$; now if additional unlabelled data $D_2 = \{\mathbf{x}\}$ is available, it is likely to provide useful information, if we assume some sort of density over $\mathbf{x}$. The reason that we are reluctant to specify such a density over $\mathbf{x}$ in speech recognition, however, is because that is where we have come from — discriminative training is used instead of full probabilistic modelling because it obtains better performance from poor models. Ideally we would like to be able to develop superior word models but what if we are stuck with a particular model space, either because of computational constraints or because of lack of creativity? Combining discriminative training with unlabelled data seems to me to be one of the current frontiers of Bayesian methods.

Other applications

The concepts of Bayesian data modelling described in this thesis have great generality and should be relevant to any experimental scientist. Example applications include the following:

Speech recognition: automated control of hidden Markov model structure
In speech recognition, selection between alternative models for a single word could be made using the evidence, and the concept of moderation (*i.e.* incorporation of error bar information) is expected to be useful when fitting a model to utterances.

Point source image reconstruction
When estimating point sources in an astronomical image, Occam's razor is needed to avoid fitting too many stars to the image.

Neurophysiology: multi–neuron recording
The task is to infer the activities of multiple neurons in a piece of brain tissue from the signals in an array of recording electrodes. This will require development of non–parametric Bayesian methods.

Density estimation
The evidence could be evaluated for the problem of choosing the number of gaussians in a mixture model, and the problem of choosing between gaussian models and more 'robust' clustering models.

Further applications in neural networks
It remains to be investigated whether these methods scale up to real, larger problems. Also this framework has yet to be applied to more sophisticated regularisers such as the mixture decay models of Hinton and Nowlan [47].

The power and unifying perspective of Bayesian methods are becoming more widely appreciated. This thesis has demonstrated their utility for adaptive models such as neural networks. There are thousands more 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.