SELECTING AMONGST LARGE CLASSES OF MODELS

B. D. RIPLEY
Department of Statistics,
University of Oxford,
Oxford OX1 3TG, UK
E-mail: ripley@stats.ox.ac.uk

Traditionally a 'model' is a family of probability distributions for the observed data parametrized by a set of parameters (of fixed and finite dimension), but it is often helpful to consider all the models considered as subsets of one model, as well as some even larger models used in 'over-fitting' as part of the validation process. Traditional distinctions between 'parametric' and 'non-parametric' models are often moot, when people now (attempt to) fit neural networks with half a million parameters. We consider how to select in such model classes.

Statisticians and other users of statistical methods have been choosing models for a long time, but the current availability of large amounts of data and of computational resources means that model choice is now being done on a scale which was not dreamt of 25 years ago. Unfortunately, the practical issues are probably less widely appreciated than they used to be, as statistical software has made it so much easier for the end user to trawl through literally thousands of models (and in some cases many more).

Model choice is a large subject, and this paper deliberately chooses to look at only some aspects of it, most particularly some of the misunderstandings about formal methods such as AIC and cross-validation. Whole books have been written about these and other aspects: two recent ones are Harrell (2001) and Burnham and Anderson (2002).

1. Why do we want to select a model?

I have slowly come to realize that this is an important question and one that is asked too seldom. First we need to ask *where do our models come from?*

- Sometimes a set of models is provided based on subject-matter theory. In my experience good theory is very rare. Sometimes

these are called *mechanistic* models. One example is the Black–Scholes theory of option pricing, which is derived from a theory and has been shown to be a good approximation, but not so good that practically important improvements cannot be made.

- Most often some simple restrictions are placed on the behaviour we expect to find, for example linear models, $AR(p)$ processes, factorial models with limited interactions. These are sometimes called *empirical* models. Note that these can be very broad classes if transformations of variables (on both sides) are allowed.
- We now have model classes that can approximate any reasonable model, for example neural networks (Ripley, 1996). Nowadays we may have the data and the computational resources to fit such models, if not necessarily the understanding to fit them well.

The main distinction I would draw is between *explanation* and *prediction*. Generally with the mechanistic models we are concerned with explaining how the world works, even though the philosophy of science teaches that we test models by their ability to predict. The third class of models is unambiguously designed to give good predictions.

For the second class, we might be doing either. When people first started to do agricultural experiments they were (it seems) both trying to find out which factors had an effect, and for those that did, how large the effect was. Nowadays many experiments are done with *microarrays* to find out which few (out of thousands) of genes are expressed differently in different experimental conditions. But regression and time-series models are most commonly used for their predictive abilities.

For *explanation*, Occam's razor applies and we want

> an explanation that is as simple as possible, but no simpler
>
> attrib Einstein

and we do have a concept of a 'true' model, or at least a model that is a good working approximation to the truth. Simplicity helps both with communicating the concepts embodied in the model and in what psychologists call *generalization*, the ability to 'work' in scenarios very different from those in which the model was studied.

On the other hand, *prediction* is like doing engineering development. All that matters is that it works, and if the aim is prediction, model selection should be based on the quality of the predictions. Workers in pattern recognition have long recognised this, and used *validation sets* to choose

between models, and *test sets* to assess the quality of the predictions from the chosen model. Because the model may be used in scenarios very different from those in which it was tested, generalization is still important, and *other things being equal* a mechanistic model or a simple empirical model has more chance of reflecting the data-generation mechanism and so of generalizing. But other things rarely *are* equal.

We should ask why we do want to *choose* a model. It does seem a widespread misconception that model choice is about 'choosing the best model'. For explanation we ought to be alert to be possibility of there being several (roughly) equally good explanatory models: when I was a young Lecturer at Imperial College I learnt this from David Cox, having already done a lot of informal model choice in applied problems in which I would have benefited from offering several alternative solutions.

For prediction I find a good analogy is that of choosing between expert opinions: if you have access to a large panel of experts, how would you use their opinions? (See Cooke, 1991.) People do tend to pick one expert and listen to him/her, but it would seem better to seek a consensus view, which translates to *model averaging* rather than model choice. Our analogy is with experts, which implies some prior selection of people with a track record: one related statistical idea is the *Occam's window* (Madigan and Raftery, 1994) which keeps only models with a reasonable record.

A major reason to choose a model appears still to be computational cost, a viewpoint of Geisser (1993, §4.1). This has become less relevant, and we discuss model averaging in a later section. Note, though, that taking a consensus view only helps sometimes with generalization. For example, Draper (1995, p. 48) has a graph of predictions of oil prices for 1981–90 made in 1980. The analysts were all confident, differed considerably from each other, and were all way off! Almost all of the uncertainty is in the 'correct' model for oil price movements, and the analysts' models all seem to be incorrect as prices went down when all the analysts predicted them to rise.

Ein-Dor and Feldmesser (1987) provide an example of the confusion between explanation and prediction that is one of my favourite teaching examples. The title says they give a *relative performance prediction model*, yet they select[a] a subset of transformed variables in seeking an explanation.

[a] the paper is a wonderful example of how **not** to do that, too.

2. A historical perspective

Let us look back 25 years to when I started to learn about this area. The set of models one could consider was severely limited by computational constraints, although packages such as GLIM 3.77 were becoming available.

Stepwise selection was the main formal tool, using hypothesis tests between a pair of nested models, e.g. F tests for regressions. Almost no one did enough tests to worry much about issues of multiple comparison. Nowadays people do tens of thousands of significance tests (see Marchini and Ripley, 2000 for an example of mine).

Residual plots were used to help assess the fit of models, but they were crude plots and limited to small datasets.

There was very little attempt to deal with choosing between models that were genuinely different explanations: Cox (1961, 1962)'s 'tests of separate families of hypotheses' existed but were little known and less used.

Formal methods of model choice were becoming available. Schwarz (1978) had proposed a criterion sometimes called SBC or BIC, although it seems to be due to Jeffreys in the 1930's. Papers by Allen (1971, 1974) and Akaike (1973, 1974) had introduced PRESS and AIC (Akaike's An Information Criterion) respectively. Cross-validation goes back at least as far as Mosteller and Wallace (1963), and Stone (1974) has been read to the Royal Statistical Society, to a less than appreciative audience.

Perhaps the only formal criterion that was in common use was Mallows' C_p criterion for regression, which I am told was well known long before Mallows' first publication (Mallows, 1973).

My impression is that these developments were held back by the lack of computational resources to try out large classes of models, and by the lack of large datasets to present challenging problems.

3. Cross-validation

Cross-validation is a much misunderstood topic in the neural networks and machine learning community.

The idea is of *leave-one-out CV* is that given a dataset of N points, we use our model-building procedure on each subset of size $N-1$, and predict the point we left out. Then the set of predictions can be summarized by some measure of prediction accuracy. Allen's PRESS (prediction sum-of-squares) used this to choose a set of variables in linear regression. Stone (1974) and Geisser (1975) pointed out we could apply this to many aspects of model choice, including parameter estimation. It is often confused with

jackknifing *a la* Quenouille and Tukey.

Having to do model-building N times can be prohibitive unless there are computational shortcuts.

In V-fold cross-validation we divide the data into V sets, amalgamate $V-1$ of them, build a model and predict the result for the remaining set. Do this V times leaving a different set out each time. How big should V be? We want the model-building problem to be realistic, so want to leave out a small proportion. We do not want too much work, so usually V is 3–10. One early advocate of this was Breiman *et al.* (1984).

Leave-one-out CV does not work well in general, as it makes too small changes to the fit. Ten-fold CV often works well, but sometimes the result is very sensitive to the partitioning used, and it is often better for comparisons than for absolute values of performance. How prediction accuracy is measured can be critical. We can now afford to average the results over several random partitions.

Stone (1974, pp. 126–7) mentioned the idea of using cross-validation not to choose between models but to combine them. This has been developed by Wolpert (1992) under the name of *stacked generalization*.

4. AIC, BIC and all that

Akaike (1973, 1974) introduced a criterion for model adequacy, first for time-series models and then more generally. He relates how his secretary suggested he call it 'An Information Criterion', AIC. Two books largely about this criterion are Sakamoto *et al.* (1986) and Burnham and Anderson (2002).

This has a very appealing simplicity:

$$AIC = -2\log(\text{maximized likelihood}) + 2p$$

where p is the number of estimated parameters. Choose the model with the smallest AIC (and perhaps retain all models within 2 of the minimum). (Despite the simplicity, quite a few people have managed to get it wrong, for example the `step` function in S-PLUS.) This is similar to Mallows' C_p criterion for regression,

$$C_p = \text{RSS}/\sigma^2 + 2p - N$$

and is the same if σ^2 is known. Both are of the form

$$\text{measure of fit} + \text{complexity penalty}$$

Schwarz's criterion, often called BIC, replaces 2 by $log\, n$ for a suitable definition of n, the size of the dataset. In the original regression context this is just the number of cases.

Derivation of AIC

Suppose we have a dataset of size N, and we fit a model to it by maximum likelihood, and measure the fit by the *deviance* D (constant minus twice maximized log-likelihood). Suppose we have m (finite) nested models.

Hypothetically, suppose we have another dataset of the same size, and we compute the deviance D^* for that dataset *at the MLE for the first dataset*. We would expect that D^* would be bigger than D, on average. In between would be the value D' obtained if we had evaluated the deviance at the true parameter values. Some Taylor-series expansions (e.g. Ripley, 1996, pp. 31–4) show that

$$E\, D^* - E\, D' \approx p, \qquad E\, D' - E\, D \approx p$$

and hence $AIC = D + 2p$ is (to this order) an unbiased estimator of $E\, D^*$. The latter is a reasonable measure of performance, the Kullback-Leibler divergence between the true model and the plug-in model (at the MLE).

These expectations are over the dataset under the assumed model.

Crucial assumptions

The assumptions needed for this argument are much less well known than they should be, and AIC is often proposed (and used) to select between m very different models.

(1) The model is true! Suppose we use this to select the order of an $AR(p)$ model. If the data really came from an $AR(p_0)$ model, all models with $p \geq p_0$ are true, but those with $p < p_0$ are not even approximately true.

    This assumption can be relaxed. Takeuchi (1976) did so, and his result has been rediscovered by Stone (1977) and many times since. However, p gets replaced by a much more complicated formula that is not simple to measure.

(2) The models are nested[b] – AIC is widely used when they are not.

[b] see the bottom of page 615 in the reprint of Akaike (1973).

(3) Fitting is by maximum likelihood. Nowadays many models are fitted by penalized methods or Bayesian averaging .... That can be worked through too, in NIC (Murata *et al.*, 1991, 1993, 1994) or p_{eff} (Moody, 1991, 1992).
(4) The Taylor-series approximations are adequate. People have tried various refinements, notably AICC (or AIC_c) given by

$$AICC = D + 2p\Big(\frac{N}{N-p+1}\Big)$$

Also, the MLEs need to be in the interior of the parameter space, even when a simpler or alternative model is true. (This is not likely to be true for variance components for example.)
(5) AIC is a reasonably good estimator of $E\,D^*$, or at least that differences between models in AIC are reasonably good estimators of differences in $E\,D^*$. This seems the Achilles' heel of AIC. For N independent samples we expect $AIC = O_p(N)$ but the variability as an estimate is $O_p(\sqrt{N})$. This reduces to $O_p(1)$ for differences between models *provided they are nested*.

AIC has been criticised in asymptotic studies and simulation studies for tending to over-fit, that is choose a model at least as large as the true model. That is a virtue, not a deficiency: this is a prediction-based criterion, not an explanation-based one.

AIC is asymptotically equivalent to leave-one-out CV for independent identically distributed samples using deviance as the loss function (Stone, 1977), and in fact even when the model is not true NIC is equivalent (Ripley, 1996).

5. Bayesian Approaches

Note the plural — I think Bayesians are rarely Bayesian in their model choices. Assume m (finite) models, exactly one of which is true.

In the Bayesian formulation (Bernardo and Smith, 1994; Draper, 1995), models are compared via $P\{M \mid \mathcal{T}\}$, the posterior probability assigned to model M given the dataset $\mathcal{T}$.

$$P\{M \mid \mathcal{T}\} \propto p(\mathcal{T} \mid M)p_M,$$

$$p(\mathcal{T} \mid M) = \int p(\mathcal{T} \mid M, \theta)p(\theta)\,\mathrm{d}\theta$$

so the ratio in comparing models M_1 and M_2 is proportional to $p(\mathcal{T} \mid M_2)/p(\mathcal{T} \mid M_1)$, known as the *Bayes factor*.

However, a formal Bayesian approach then averages predictions from models, weighting by $P\{M \mid \mathcal{T}\}$, unless a very peculiar loss function is in use.

Suppose we just use the Bayes factor as a guide. The difficulty is in evaluating $p(\mathcal{T} \mid M)$. Asymptotics are not useful for Bayesian methods, as the prior on θ is often very important in providing smoothing, yet asymptotically negligible. One approximation is to take $\widehat{\theta}$ as the mode of the posterior density and V as the inverse of the Hessian of $-\log p(\widehat{\theta} \mid \mathcal{T})$ (since for a normal density this is the covariance matrix); we can hope to find $\widehat{\theta}$ and V from the maximization of

$$\log p(\theta \mid \mathcal{T}) = L(\theta; \mathcal{T}) + \log p(\theta) + \text{const}$$

Let $E(\theta) = -L(\theta; \mathcal{T}) - \log p(\theta)$, so this has its minimum at $\widehat{\theta}$ and Hessian there of V^{-1}.

$$\begin{aligned} p(\mathcal{T} \mid M) &= \int p(\mathcal{T} \mid \theta)\, p(\theta)\, \mathrm{d}\theta = \int \exp -E(\theta)\, \mathrm{d}\theta \\ &\approx \exp -E(\widehat{\theta}) \int \exp[-\tfrac{1}{2}(\theta - \widehat{\theta})^T V^{-1} (\theta - \widehat{\theta})]\, \mathrm{d}\theta \\ &= \exp -E(\widehat{\theta})\ (2\pi)^{p/2} |V|^{1/2} \end{aligned}$$

via a Laplace approximation to the integral. Thus

$$\log p(\mathcal{T} \mid M) \approx L(\widehat{\theta}; \mathcal{T}) + \log p(\widehat{\theta}) + \tfrac{p}{2} \log 2\pi + \tfrac{1}{2} \log |V|.$$

It may be feasible to use this directly for model selection.

If we suppose θ has a prior which we may approximate by $N(\theta_0, V_0)$, we have

$$\begin{aligned} \log p(\mathcal{T} \mid M) \approx L(\widehat{\theta}; \mathcal{T}) &- \tfrac{1}{2}(\widehat{\theta} - \theta_0)^T V_0^{-1} (\widehat{\theta} - \theta_0) \\ &- \tfrac{1}{2} \log |V_0| + \tfrac{1}{2} \log |V| \end{aligned}$$

and V^{-1} is the sum of V_0^{-1} and the Hessian H of the log-likelihood at $\widehat{\theta}$. Thus

$$\log p(\mathcal{T} \mid M) \approx L(\widehat{\theta}; \mathcal{T}) - \tfrac{1}{2}(\widehat{\theta} - \theta_0)^T V_0^{-1} (\widehat{\theta} - \theta_0) - \tfrac{1}{2} \log |H|.$$

If we assume that the prior is very diffuse we can neglect the second term, so the penalty on the log-likelihood is $-\frac{1}{2} \log |H|$.

For a random sample of size n from the assumed model, this might be roughly proportional to $-(\frac{1}{2} \log n)\, p$ provided the parameters are identifiable. This is the proposal of Schwarz (1978).

Crucial assumptions

(1) The data were generated as an independent, identically distributed random sample, and originally for linear models only. It is not clear what n should be for, say, a random effects model.
(2) Choosing a single model is relevant in the Bayesian approach.
(3) The model is true.
(4) The prior can be neglected. We may not obtain much information about parameters which are rarely effective, even in very large samples.
(5) The simple asymptotics are adequate and that the rate of data collection on each parameter would be the same. We should be interested in comparing different models for the same n, and in many problems p will be comparable with n.

Note that as this is trying to choose an explanation, we would expect BIC to neither overfit nor underfit, and there is some theoretical support for that.

6. Deviance Information Criterion

Named by Spiegelhalter *et al.* (2002) in a Bayesian setting where prior information is not negligible, and the model is assumed to be a good approximation but not necessarily true.

In GLMs (and elsewhere) the *deviance* is the difference in twice maximized log likelihood between the *saturated* model and the fitted model, or

$$D(\theta) = \text{deviance}(\theta) = \text{const}(\mathcal{T}) - 2L(\theta; \mathcal{T})$$

and in GLMs we use $D(\widehat{\theta})$ as the (scaled) (residual) deviance.

Define

$$p_D = \overline{D(\theta)} - D(\overline{\theta})$$

The first overline means averaging θ over $p(\theta \,|\, \mathcal{T})$, and the second means our estimate of the 'least false' parameter value, usually the posterior mean of θ (but perhaps the median or mode of the posterior distribution). Then define

$$DIC = D(\overline{\theta}) + 2\, p_D$$

Clearly DIC is AIC-like, but

- Like NIC it allows for non-ML fitting, in particular for the regularization effect of the prior that should reduce the effective number of parameters.
- It is not necessary (but is usual) that $p_D \geq 0$.
- DIC is explicitly meant to apply to non-nested non-IID problems.
- DIC is intended to be approximated via MCMC samples from the posterior density of θ given $\mathcal{T}$. On the other hand, DIC needs an explicit formula for the likelihood (up to a model-independent normalizing constant).

7. Model Averaging

For prediction purposes (and that applies to almost all Bayesians) we should average the predictions over models. What do we average?

The probability predictions made by the models.

For linear regression this amounts to averaging the coefficients over the models (being zero where a regressor is excluded), and this becomes a form of shrinkage. Other forms of shrinkage like ridge regression may be as good at very much lower computational cost.

Note that we may not want to average over all models. We may want to choose a subset for computational reasons, or for plausibility.

How do we choose the weights?

In the Bayesian theory this is clear, via the Bayes factors. In practice this is discredited. Even if we can compute them accurately (and via Markov Chain Monte Carlo we may have a chance), we assume that one and exactly one model is true. In practice Bayes factors can depend on aspects of model inadequacy which are of no interest. I first encountered that in Ripley (1992), where we fitted formal probability models to images (and therefore had tens of thousands of observations). There was a common noise model but different priors for the different models. We were able to calculate Bayes factors approximately by MCMC in a week or so, and we pleased to see that that the factors were very decisive. After some checking, we discovered that they were very decisively picking the wrong model. There was a 'true' model (the models represented different species of nematodes) but a lot of investigation showed that the 'noise' model was interacting with the texture of the nematodes.

Alternative approaches are via cross-validation (goes back to Stone, 1974) and via bootstrapping (LeBlanc and Tibshirani, 1993). This can also be viewed as an extended estimation problem, with the weights depending on the sample via a model (e.g. a multiple logistic); so-called *stacked generalization* (Wolpert, 1992) and *mixtures of experts* (Jacobs *et al.*, 1991).

Bagging, boosting, random forests

Model averaging ideas have been much explored in the field of classification trees.

In *bagging* (Breiman, 1996a,b) models are fitted to bootstrap resamples of the data, and weighted equally. Breiman (1996b) motivates this for *unstable* methods such as classification trees in which a small change in the training set can lead to a large change in the classifier. A variant on this idea which has been suggested many times is to add 'noise' to the training set, randomly perturbing either the feature vectors $\mathbf{x}$ or the classes c (or both). Further along this line, we could model the joint distribution of $(\mathbf{X}, C)$ and create new training sets from this distribution. Bagging can be seen as the rather extreme form of this procedure in which the model is the empirical distribution. (Krogh and Vedelsby, 1995, use cross-validation rather than re-sampling, and consider designing training sets weighted towards areas where the existing classifiers are prone to disagree.)

In *boosting* (Schapire, 1990; Freund, 1990; Drucker *et al.*, 1994; Freund, 1995; Freund and Schapire, 1995) each additional model is chosen to (attempt to) repair the inadequacies of the current averaged model by resampling biased towards the mistakes. The idea is to *design* a series of training sets and use a combination of classifiers trained on these sets. (Majority voting and linear combinations have both been used.) There have been many papers on this topic, as well as empirical tests which tend to show (e.g., Quinlan, 1996) that boosting often does well but occasionally does disastrously.

In *random forests* (Breiman, 2001) the tree-construction algorithm randomly restricts itself at the choice of each split, to create a 'forest' of trees from a single training set.

8. Practical model selection in 2004

The concept of a model is much larger than it was 25 years ago. Even a decade ago, people attempted to fit neural networks with half a million free parameters. We are no longer so tied to maximum likelihood estimation,

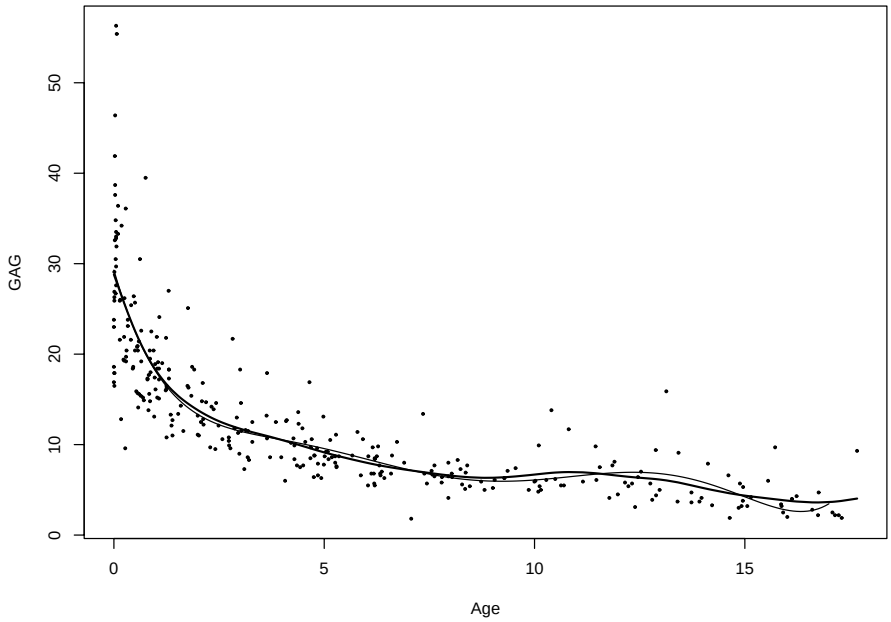

Figure 1. Two smooth curves fitted to the concentration of the chemical GAG in the urine of 314 children aged 0–18 years.

and fit models to much larger datasets. The latter almost inevitably means that we fit more complex models, and 'smooth' terms are often used in place of linear[c] terms.

Large model classes often overlap very considerably. There are many ways to obtain a smooth curve like Figure 1. The traditional approach would be to fit a polynomial, and one of the curves is a degree-six polynomial chosen by forwards stepwise selection. The other is a smoothing spline, with the degree of smoothness chosen by GCV.[d] There are many alternative approaches, including neural networks and local polynomials (Wand and Jones, 1995; Loader, 1999). These can all fit very similar curves, and the issue of choosing between the model classes is rather a moot one.

Alternative explanations with roughly equal support are commonplace: model averaging seems a good solution. Selecting several models, studying their predictions and taking a consensus is also a good idea, *when time permits* and when *non-quantitative information is available*. As Figure 1

[c] or low-order polynomial

[d] *generalized cross-validation*, which is not in fact cross-validation as defined here.

shows, we may need other information to choose between very different formulae with similar predictions, in so far as we can choose at all.

'Regression diagnostics' are often based on approximations to overfitting or case deletion. Now we can (and some of us do) fit extended models with smooth terms or use fitting algorithms that automatically downweight groups of points. (I rarely use least squares these days.) It is still all too easy to select a complex model just to account for a tiny proportion of aberrant observations.

Although we do have more tools available than at the start of my career, it seems to me that model selection has actually got harder: as we explore more of the statistical model world we encounter more and more chasms awaiting the unwary. It worries me how causally AIC and its allies are used, and hope this paper will go some way to raising awareness of the limitations of formal methods of model selection.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (Eds B. N. Petrov and F. Cáski), pp. 267–281, Budapest. Akademiai Kaidó. Reprinted in *Breakthroughs in Statistics*, eds Kotz, S. & Johnson, N. L. (1992), volume I, pp. 599–624. New York: Springer.

Akaike, H. (1974) A new look at statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Allen, D. M. (1971) Mean square error of prediction as a criterion for selecting variables. *Technometrics*, **13**, 325–331.

Allen, D. M. (1974) The relationship between variable selection and data augmentation and a method of prediction. *Technometrics*, **16**, 467–475.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. Chichester: Wiley.

Breiman, L. (1996a) Bagging predictors. *Machine Learning*, **24**, 123–140.

Breiman, L. (1996b) The heuristics of instability in model selection. *Annals of Statistics*, **24**, 2350–2383.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks/Cole.

Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference*. New York: Springer, second edition.

Cooke, R. M. (1991) *Experts in Uncertainty. Opinion and Subjective Probability in Science*. New York: Oxford University Press.

Cox, D. R. (1961) Tests of separate families of hypotheses. In *Proc. 4th Berkeley Symposium* (Ed. J. Neyman), volume 1, pp. 105–123, University of California Press. University of California Press.

Cox, D. R. (1962) Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society series B*, **24**, 406–424.

Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society series B*, **57**, 45–97.

Drucker, H., Cortes, C., Jaeckel, L. D., LeCun, Y. and Vapnik, V. (1994) Boosting and other ensemble methods. *Neural Computation*, **6**, 1289–1301.

Ein-Dor, P. and Feldmesser, J. (1987) Attributes of the performance of central processing units: A relative performance prediction model. *Communications of the ACM*, **30**, 308–317.

Freund, Y. (1990) Boosting a weak learning algorithm by majority. In *Proceedings of the Third Workshop on Computational Learning Theory*, pp. 202–216. Morgan Kaufmann.

Freund, Y. (1995) Boosting a weak learning algorithm by majority. *Information and Computation*, **121**(2), 256–285.

Freund, Y. and Schapire, R. E. (1995) A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*, pp. 23–37. Springer.

Geisser, S. (1975) The predictive sample reuse method with applications. *Journal of the American Statistical Association*, **70**, 320–328.

Geisser, S. (1993) *Predictive Inference: An Introduction*. New York: Chapman & Hall.

Harrell, Jr., F. E. (2001) *Regression Modeling Strategies, with Applications to Linear Models, Logistic Regression and Survival Analysis*. New York: Springer-Verlag.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. and Hinton, G. E. (1991) Adaptive mixtures of local experts. *Neural Computation*, **3**, 79–87.

Krogh, A. and Vedelsby, J. (1995) Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems 7. Proceedings of the 1994 Conference* (Eds G. Tesauro, D. S. Touretzky and T. K. Leen), pp. 231–238, Cambridge, MA. MIT Press.

LeBlanc, M. and Tibshirani, R. J. (1993) Combining estimates in regression and classification. Preprint, Depts of Preventive Medicine and Biostatis-

tics and of Statistics, University of Toronto.

Loader, C. (1999) *Local Regression and Likelihood*. New York: Springer-Verlag.

Madigan, D. and Raftery, A. E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the Royal Statistical Society*, **89**, 1535–1546.

Mallows, C. L. (1973) Some comments on C_p. *Technometrics*, **15**, 661–675.

Marchini, J. L. and Ripley, B. D. (2000) A new statistical approach to detecting significant activation in functional MRI. *NeuroImage*, **12**, 366–380.

Moody, J. E. (1991) Note on generalization, regularization and architecture selection in nonlinear learning systems. In *First IEEE-SP Workshop on Neural Networks in Signal Processing*, pp. 1–10, Los Alamitos, CA. IEEE Computer Society Press.

Moody, J. E. (1992) The *effective* number of parameters: an analysis of generalization and regularization in nonlinear learning systems. In *Advances in Neural Information Processing Systems 4. Proceedings of the 1991 Conference* (Eds J. E. Moody, S. J. Hanson and R. P. Lippmann), pp. 847–854, San Mateo, CA. Morgan Kaufmann.

Mosteller, F. and Wallace, D. L. (1963) Inference in an authorship problem. *Journal of the American Statistical Association*, **58**, 275–309.

Murata, N., Yoshizawa, S. and Amari, S. (1991) A criterion for determining the number of parameters in an artificial neural network model. In *Artificial Neural Networks. Proceedings of ICANN-91* (Eds T. Kohonen, K. Mäkisara, O. Simula and J. Kangas), volume I, pp. 9–14, Amsterdam. North Holland.

Murata, N., Yoshizawa, S. and Amari, S. (1993) Learning curves, model selection and complexity of neural networks. In *Advances in Neural Information Processing Systems 5. Proceedings of the 1992 Conference* (Eds S. J. Hanson, J. D. Cowan and C. L. Giles), pp. 607–614, San Mateo, CA. Morgan Kaufmann.

Murata, N., Yoshizawa, S. and Amari, S. (1994) Network information criterion—determining the number of hidden units for artificial neural network models. *IEEE Transactions on Neural Networks*, **5**, 865–872.

Quinlan, J. R. (1996) Bagging, boosting, and C4.5. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, Menlo Park, CA. AAAI Press.

Ripley, B. D. (1992) Classification and clustering in spatial and image data. In *Analyzing and Modeling Data and Knowledge* (Ed. M. Schader), pp.

93–105, Berlin. Springer.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1986) *Akaike Information Theory Statistics*. Dordrecht: Reidel.

Schapire, R. E. (1990) The strength of weak learnability. *Machine Learning*, **5**(2), 197–227.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society series B*, **64**, 583–639.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society series B*, **36**, 111–147.

Stone, M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society series B*, **39**, 44–47.

Takeuchi, K. (1976) Distribution of informational statistics and a criterion of fitting. *Suri-Kagaku*, **153**, 12–18. [In Japanese].

Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*. London: Chapman & Hall.

Wolpert, D. H. (1992) Stacked generalization. *Neural Networks*, **5**, 241–259.