Model Choice

Brian D. Ripley

Professor of Applied Statistics
University of Oxford

`ripley@stats.ox.ac.uk`
`http://stats.ox.ac.uk/~ripley`

Sub-titles

- Towards general principles for model selection.
- 'Give up your small ambitions'
- Check your assumptions –
  - of the models
  - of the model selection theories

Model Choice in 1977

That's both 25 years ago and when I started to learn about this.

- 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.

  No one did enough tests to worry much about multiple comparisons issues.
- Residual plots were used, 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's (1961) 'tests of separate families of hypotheses' existed but was little known and less used.

But the world was changing . . . .

Why do we want to choose a model?

It took me a long while to realize how profound a question that was.

Explanation *vs* Prediction

This causes a lot of confusion. 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, for

> all models are false, but some are useful
>
> G.E.P. Box, 1976

Explanation is like doing scientific research.

On the other hand, prediction is like doing engineering development. All that matters is that it works. And if the aim is prediction, model choice 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.

One of my favourite teaching examples is

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

which despite its title selects a subset of transformed variables. The paper is a wonderful example of how **not** to do that, too.

Where do the models come from?

Not unrelated to the above.

- Sometimes a set of models is provided based on subject-matter theory. In my experience good theory is very rare. Sometimes called *mechanistic* models. One example is the Black–Scholes theory of option pricing.
- 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. 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. And we may have subsets within these such as (generalized) additive models. Nowadays we may have the data and the computational resources to fit such models.

Two 1980's–style Examples

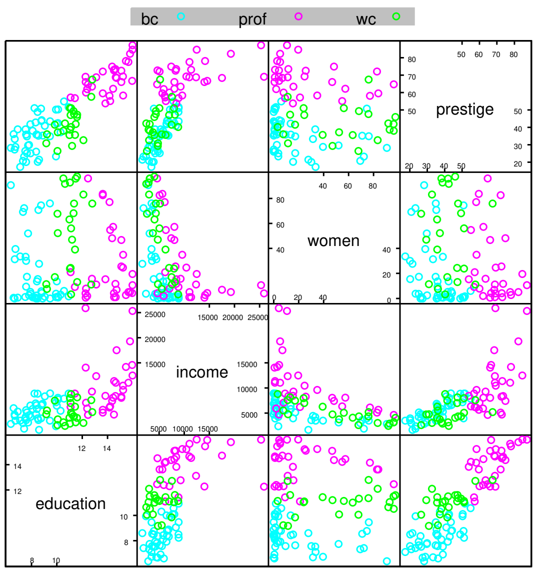
From

Fox, J. A. (2002) *An R and S-PLUS Companion to Applied Regression.* Sage.

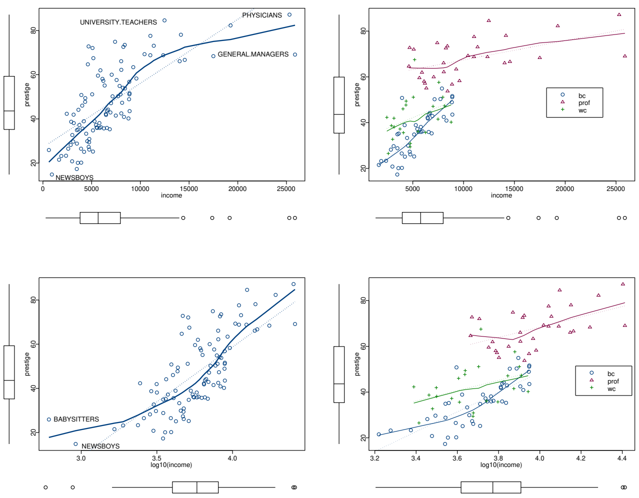
Prestigious occupations

Data from a Canadian study of 102 occupations. The response is scores of 'prestige' in an opinion survey. Possible explanatory values are 'income' (the average income), 'education' (the average number of years of education), 'women', the proportion of women and 'class', a categorical variable with levels for professional/managerial, white collar or blue collar.

A scatterplot matrix suggests that we want to transform income.

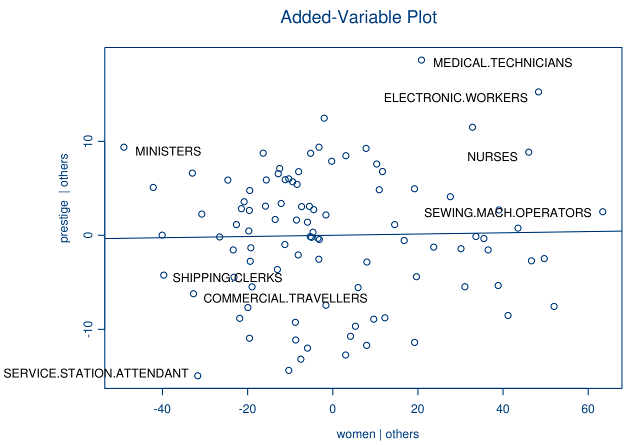


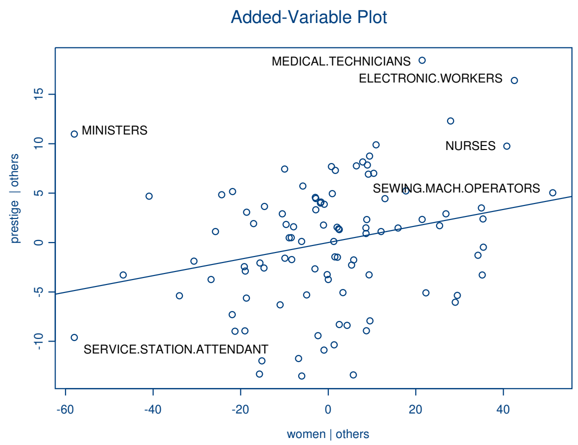
Scatterplot matrix of Canadian occupations.




```
> fit1 <- lm(prestige ~ income + education +
             women + type)
> summary(fit1)

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) -0.814  5.331     -0.153   0.879
     income  0.001  0.000      3.976   0.000
  education  3.662  0.646      5.671   0.000
      women  0.006  0.030      0.212   0.832
   typeprof  5.905  3.938      1.500   0.137
     typewc -2.917  2.665     -1.094   0.277
```

```
Residual standard error: 7.13 on 92 degrees of freedom
Multiple R-Squared: 0.835

> fit2 <- lm(prestige ~ log10(income) + education +
             women + type)
> summary(fit2)

Coefficients:
                 Value Std. Error  t value Pr(>|t|)
  (Intercept) -115.672  18.802     -6.152    0.000
log10(income)   33.745   5.322      6.340    0.000
    education    2.974   0.602      4.940    0.000
        women    0.084   0.032      2.601    0.011
     typeprof    5.292   3.556      1.488    0.140
       typewc   -3.216   2.407     -1.336    0.185

Residual standard error: 6.44 on 92 degrees of freedom
Multiple R-Squared: 0.865
```

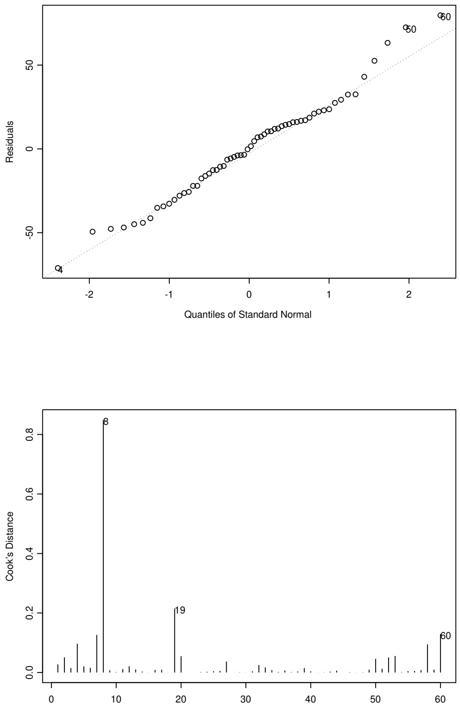


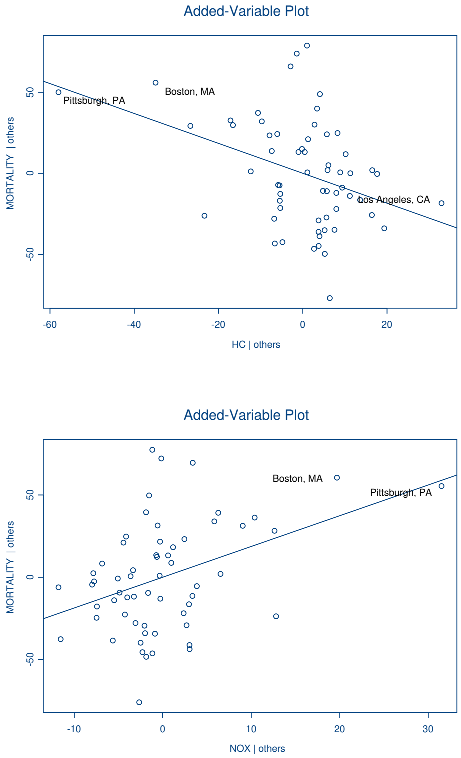


The Effects of Pollution on Mortality

Data on the mortality rates in 60 US cities, with possible explanatory variables 4 weather variables, 8 census variables and pollution levels of hydrocarbons, NO_x and SO_2.

```
> fit <- lm(MORTALITY ~ ., data=mortality)
> fit2 <- stepAIC(fit, direction ="both", trace=F)
> fit2$anova
        Step Df Deviance Resid. Df Resid. Dev AIC
1                              44      53631 440
2     - POOR  1        3       45      53634 438
3 - HUMIDITY  1       13       46      53647 436
4 - WHITECOL  1       18       47      53664 434
5    - SOUND  1      215       48      53879 432
6      - SO2  1      295       49      54175 430
7  - DENSITY  1     1168       50      55342 430

> dropterm(fit2, test="F")
         Df Sum of Sq   RSS AIC F Value Pr(F)
  <none>               55342 430
  PRECIP  1      5443 60785 433     4.9 0.031
 JANTEMP  1     11709 67052 439    10.6 0.002
JULYTEMP  1      6251 61593 434     5.6 0.021
  OVER65  1      2590 57932 430     2.3 0.132
   HOUSE  1      5849 61191 434     5.3 0.026
    EDUC  1     12175 67518 440    11.0 0.002
NONWHITE  1     25631 80973 450    23.2 0.000
      HC  1      9238 64580 437     8.3 0.006
     NOX  1     10433 65776 438     9.4 0.003
```





Cross-validation

A much misunderstood topic!

Leave-one-out CV

The idea 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. Idea goes back at least as far as Mosteller & Wallace (1963).

Allen's (1971, 4) PRESS (prediction sum-of-squares) used this to choose a set of variables in linear regression.

Stone (1974) / Geisser (1975) pointed out we could apply this to many aspects of model choice, including parameter estimation.

This is *not* jackknifing *a la* Quenouille and Tukey.

Having to do model-building N times can be prohibitive unless there are computational shortcuts (as there for linear regression, LDA, QDA and smoothing splines).

V-fold cross-validation

Divide the data into V sets, and 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 don't want too much work. So usually V is 3–10.

One early advocate of this was CART (the book).

Does it work?

Leave-one-out CV does not work well in general. It makes too small changes to the fit.

10-fold CV often works well, but sometimes the result is very sensitive to the partitioning used. Often better for comparisons than for absolute values of performance.

How prediction accuracy is measured can be critical.

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.

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 that, quite a few people have managed to get it wrong!

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. (This was first published by Mallows in 1973 but is much older.)

Both are of the form

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

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' if we had evaluated the deviance at the true parameter values. Some Taylor-series expansions 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 $E\,D^*$. And that 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

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. p gets replaced by a much more complicated formula.
2. The models are nested – AIC is widely used when they are not.
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 or Moody's p_{eff}.
4. The Taylor-series approximations are adequate. People have tried various refinements, notably AICC (or AIC$_c$) given by

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

   Also, the MLEs need to be in the interior of the parameter space, even when a simpler or alternative model is true. (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. $AIC = O_p(N)$ 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 iid samples and using deviance as the loss function (Stone, 1977), and in fact even when the model is not true NIC is equivalent.

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, models are compared via $P\{M \mid \mathcal{T}\}$, the posterior probability assigned to model M.

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

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

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. And this has been used for a long time, despite recent attempts to claim the credit for 'Bayesian Model Averaging'.

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 choice.

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), sometimes called SBC or BIC (although it seems to be due to Harold Jeffreys in the 1930's).

As with AIC, the model with minimal BIC is chosen.

Crucial assumptions

1. The data were derived as an iid sample. (What about e.g. random effects models?) (Originally for linear models only.)
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 it to neither overfit nor underfit, and there is some theoretical support for that.

There are other (semi-)Bayesian approaches, including DIC.

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 (unscaled) (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 $\mathcal{T}$.
- OTOH, DIC needs an explicit formula for the likelihood (up to a model-independent normalizing constant).

Model Averaging

For prediction purposes (and that applies to almost all Bayesians) we should average the predictions over models. We **do not choose** a single model.

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 MCMC 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.
- Via cross-validation (goes back to Stone, 1974).
- Via bootstrapping (LeBlanc & Tibshirani, 1993).
- As an extended estimation problem, with the weights depending on the sample via a model (e.g. a multiple logistic); so-called *stacked generalization* and *mixtures of experts*.

Bagging, boosting, random forests

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

In *bagging* models are fitted from bootstrap resamples of the data, and weighted equally.

In *boosting* each additional model is chosen to (attempt to) repair the inadequacies of the current averaged model by resampling biased towards the mistakes.

In *random forests* the tree-construction algorithm randomly restricts itself at the choice of each split.

Model Choice in 2002

- The concept of a model ought to be much, much larger than in 1977.
- Many models are not fitted by maximum likelihood, to very large datasets.
- Model classes can often overlap in quite extensive ways.
- There are lots of formal 'figures of adequacy' for a model. Some have proved quite useful, but
  - Their variability as estimators can be worrying large.
  - Computation, e.g. of 'effective number of degrees of freedom', can be difficult.
  - Their implicit measure of performance can be overly sensitive to certain aspects of the model which are not relevant to our problem.

  The assumptions of the theories need to be checked, as the criteria are used way outside their known spheres of validity (and in some cases where they are clearly not valid).

- Formal training/validation/test sets, or the cross-validatory equivalents, are a very general and safe approach.
- 'Regression diagnostics' are often based on approximations to over-fitting or case deletion. Now we can (and some of us do) fit extended models with smooth terms or use fitting algorithms that downweight groups of points. (I rarely use least squares these days,)
- 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*.
- I do use AIC quite a lot, especially in simpler problems. Hence the `stepAIC` function for S-PLUS/R!

Some 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 & 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.

Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference*, Second edition, Springer.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks.* CUP.

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. *JRSSB*, **64**, 583–639.