Model Choice

Lecture 1

Brian Ripley

`http://www.stats.ox.ac.uk/∼ripley/ModelChoice/`

Manifesto

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 selection is now being done on a scale which was not dreamt of a generation ago.

Unfortunately, the practical issues are probably less widely appreciated than they used to be, as statistical software and the advent of AIC, BIC (and so on) has made it so much easier for the end user to trawl through literally thousands of models (and in some cases many more).

Traditional distinctions between ‘parametric’ and ‘non-parametric’ models are often moot, when people now (attempt to) fit neural networks with half a million parameters.

What is a model?

Of the senses you will find in a dictionary, the most appropriate is

> *a simplified (often mathematical) description of a system etc. to assist calculations and predictions.*

$(\Omega, \mathcal{F}, P)$? The issue is what P represents:

- A true (and completely known) generating mechanism.
- Some convenient and partially known mechanism, so usually one of $(P_\theta, \theta \in \Theta)$.
- Our best understanding of the generating mechanism after parameter estimation and (perhaps) model choice.

Where do the models come from?

- 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.

What are models good for?

This is often taken for granted.

- For *explanation*, to test a possible causal mechanism.
- For *prediction*.
- For decision making (which is more or less the same as prediction).
- For adjustment.

Quite a lot of the discussion in the literature is (it seems to me) at cross-purposes because it is making different and unstated assumptions about the purpose of models.

The purpose of models is not to fit the data but to sharpen the questions.

Samuel Karlin, 1983
(11th R A Fisher Memorial Lecture)

A theory has only the alternative of being right or wrong. A model has a third possibility: it may be right, but irrelevant.

Manfred Eigen, 1973

No one trusts a model except the person who wrote it. Everyone trusts an observation except the person who made it.

Why do we want to choose a model?

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

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
>
> George 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.

More on CPUs' performance

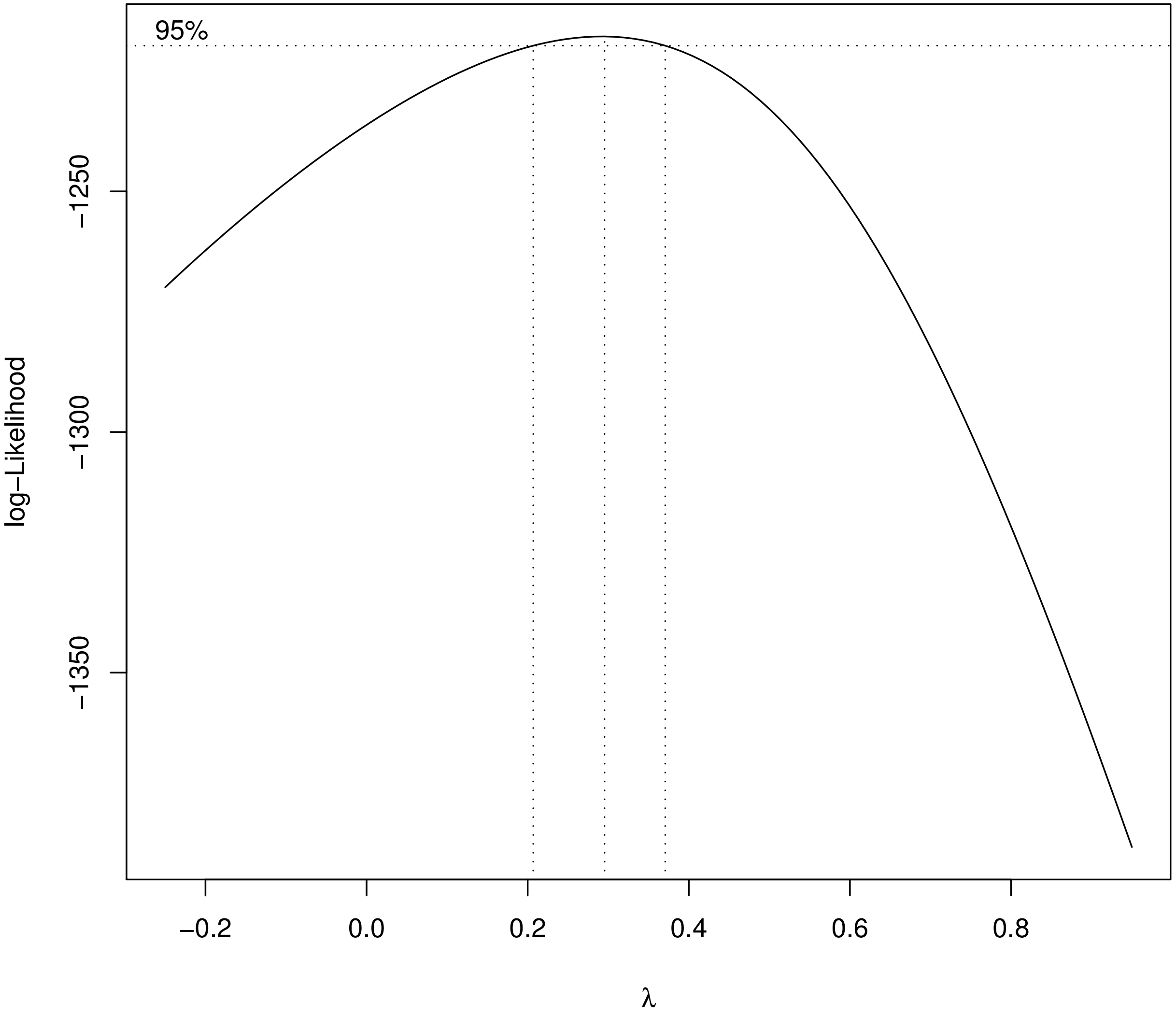
There were six machine characteristics on 209 computer 'models':

- the cycle time (nanoseconds),
- the cache size (Kb),
- the minimum and maximum possible main memory size (Kb)
- the minimum and maximum possible number of channels.

How much memory and how many channels the actual machine tested had is unspecified.

The original paper gave a linear regression for the **square root** of performance, but log scale looks more intuitive. We have a methodology to test that, from Box & Cox (1964).

Box–Cox transformations



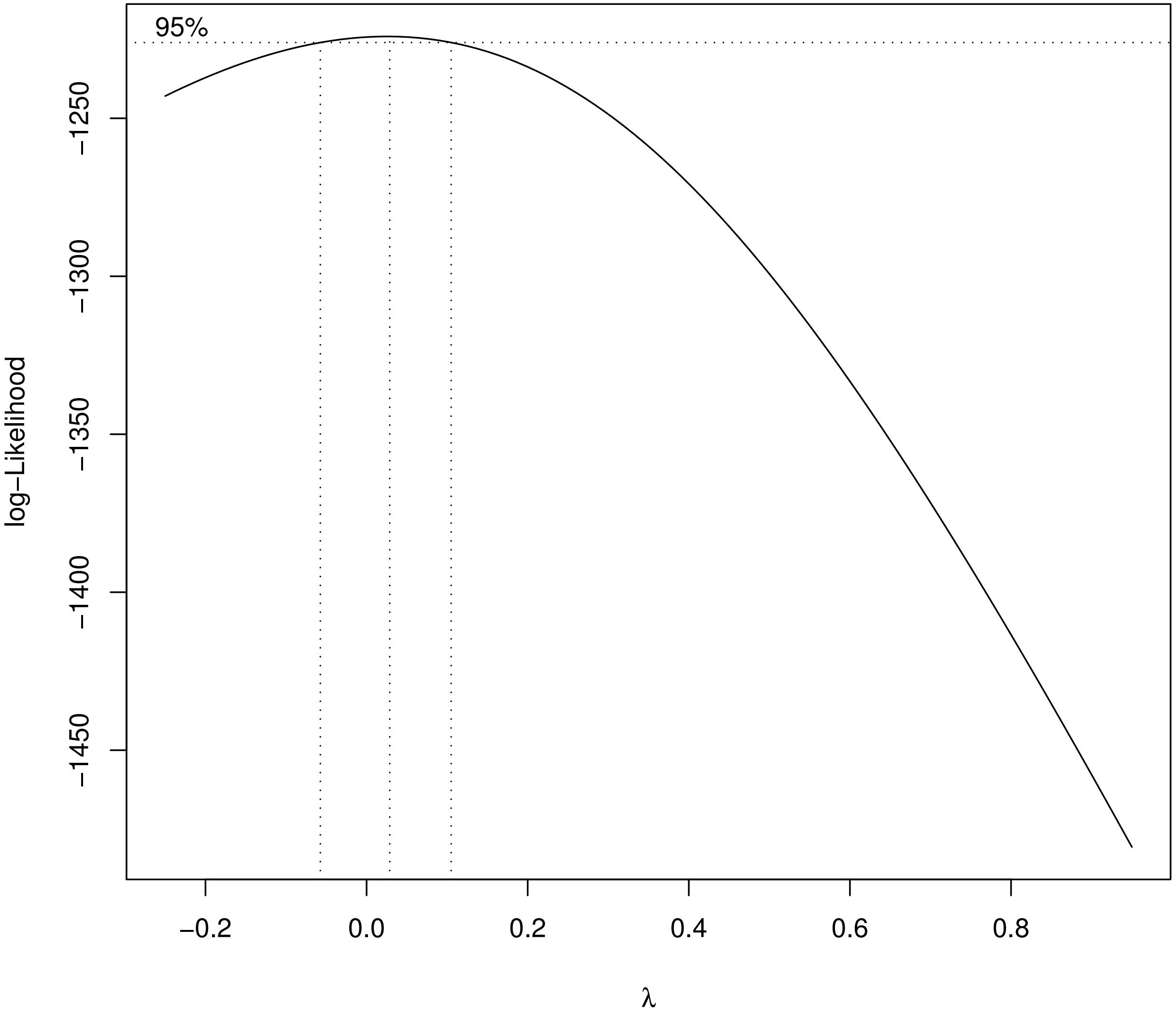
That is not what we were expecting!

Caveat: what did we transform?

We only transformed the response: it is natural to transform the regressors as well, so we need to choose several transformations simultaneously. We have technology to do that, even with non-parametric smooth functions (ACE, AVAS, . . .) but it is not very reliable.

Old-fashioned methods work: we discretized the continuous regressors into four groups and used these as categorical predictors.

Box–Cox transformations revisited



which is rather satisfying. We need a broad enough class of models.

Plug-in *vs* predictive approaches

The way we teach applied statistics we

- posit a parametric model $(P_\theta, \theta \in \Theta)$.
- estimate the parameters as $\widehat{\theta}$.
- (perhaps) do some diagnostic checks on the fit of $P_{\widehat{\theta}}$.
- (to a very large extent) act as if $P_{\widehat{\theta}}$ is the 'true' model.

In the pattern recognition literature this is known as the *plug-in* approach.

For a long time a few people objected, notably Aitchison and Geisser. They said that we should be using

$$\hat{P}(A) = \int P_\theta(A) p(\theta \,|\, \text{data}) \, \mathrm{d}\theta$$

known as the *predictive* approach.

In this approach the P of our model is our current best summary of our knowledge about the data-generating mechanism.

This does seem the right thing to use for prediction and decision-making purposes. For many simple families of models (linear regression, linear discriminant analysis) is makes only a little difference, but for others (e.g. logistic regression) it does.

Note that this is more general than it seems at first sight, as variable selection problems (for example) reduce to setting some components of θ to zero. Thus the predictive approach involves averaging over uncertainty between (narrow-sense) models as well as within parametric models.

Why select a model at all?

It does seem a widespread misconception that model choice is about

choosing the best model

For *explanation* we ought to be alert to the possibility of there being several (roughly) equally good explanatory models.

I learnt that from David Cox after having already done a lot of informal model choice in applied problems in which I would have benefited from offering several alternative solutions.

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. So there is a premium on few models.

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?

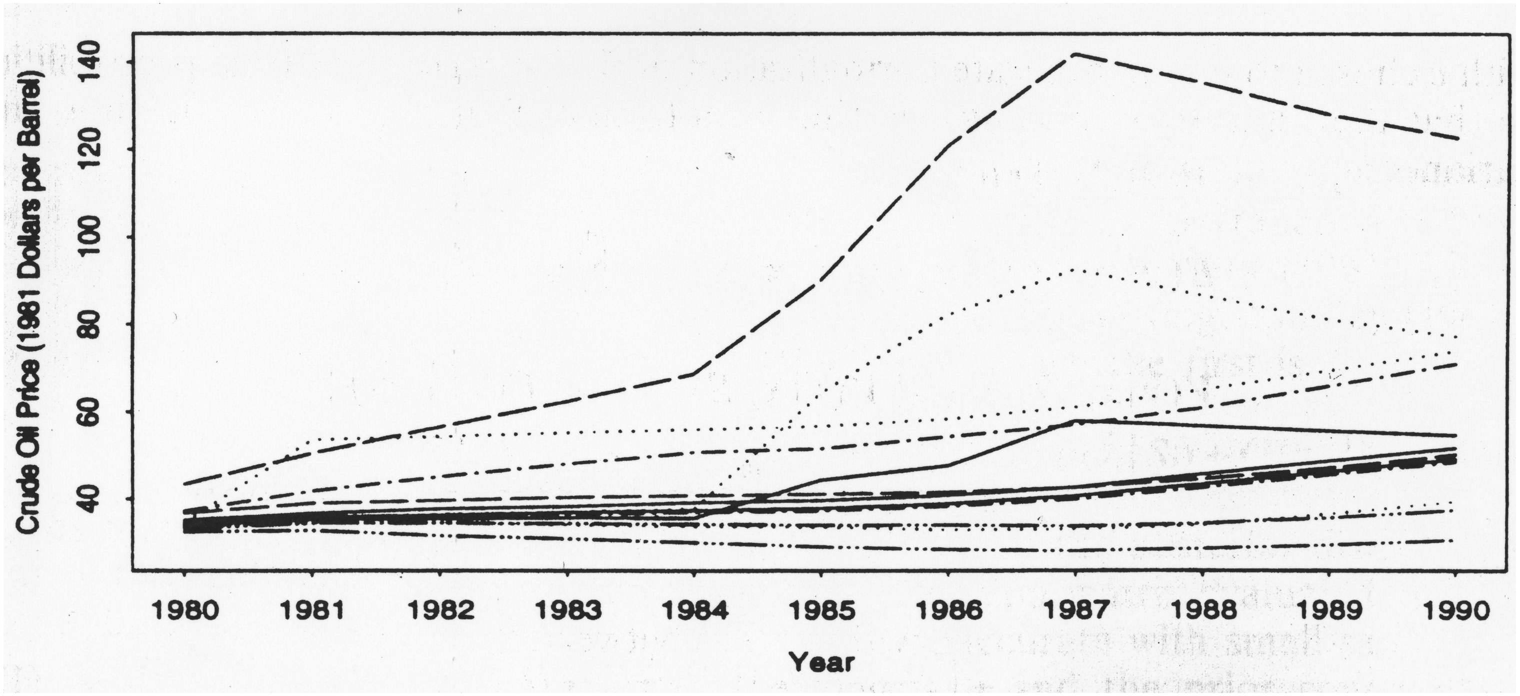
People do tend to pick one expert ('guru') and listen to him/her, but it would seem better to seek a consensus view, which translates to *model averaging* rather than model choice. Our analogy is with experts, which implies some prior selection of people with a track record: one related statistical idea is the *Occam's window* (Madigan & Raftery, 1994) which keeps only models with a reasonable record.

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.

All the models/experts may be wrong

Note that taking a consensus view only helps sometimes with generalization.

Draper (1995) 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 — the oil price was \$13 in 1986!



Computational cost

A major reason to choose a model appears still to be computational cost, a viewpoint of Geisser (1993). Even if we can fit large families of models, we may have time to consider the predictions only from a few.

> The question then is what we do once we select the best model? Presumably it affords the "best" single description amongst those entertained. Should we now use it for prediction? If we do, we know it is not optimal under any loss function except one that reflects a principle of parsimony that states one should only use one of the models for prediction.

A much-quoted example is a NIST study on reading hand-written ZIP codes, which have to be read in about 1/2 second each to be useful in a sorting machine. The best known predictions come by combining tens of models.

Models for adjustment

Here is a usage where selection is hard to defend.

The *MSc Case Studies* class has considered two problems where linear regression models were used for adjustment.

Effect of lead on blood pressure

Measurements of systolic blood pressure on over 20,300 males. Interest is in predict change in blood pressure with exposure to lead (as measured by blood samples). Stepwise selection was done on ca 40 possible adjustment variables.

Effect of punishment on crime rates

Ehrlich's data on crime rates in 47 US states, with 15 explanatory variables, 12 for adjustment and three of interest (income inequality, probability of imprisonment, length of sentence).

An historical perspective — Model choice in 1977

That's 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.

  Few people 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 ....

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), and 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.

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

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

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 the CART book (Breiman, Friedman, Olshen & Stone, 1984) and program.

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. We can ‘average’ over several random partitions.

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. Both are of the form

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

Schwarz's (1978) criterion, often called BIC or SBC, 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.

BIC was anticipated by work of Harold Jeffreys in the 1930's.

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 $\ldots$. 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

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

   Also, the MLEs need to be in the interior of the parameter space, even when a simpler or alternative model is true. (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 (Ripley, 1996).

Next week's lecture

- Other formal criteria, especially those with a Bayesian flavour.
- Selection where we have a specific goal in mind (FIC)
- Shrinking rather than dropping variables.
- Combining different models.
- The state of the art in 2010.