Tina Memo No. 2011-011
Internal.

Model Selection Using AIC and Degree of Freedom Methods.

Neil A Thacker

Last updated
4/10/2011

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Model Selection Using AIC and Degree of Freedom Methods.

N.A. Thacker 4/10/11

This document was written following a discussion with JG, regarding the question "How are AIC and more conventional χ^2/DOF methods for model order selection related?" in which he suggested a diagramatic approach for investigating the issue. This approach did seem to offers something interesting, not only in clarifying the opertation of the methods but also in supporting some form of comparison. This document summarises my interpretation and conclusions.

Introduction

We often need to analyse data d_n in order to extract parameters a_k, and this can be done by optimising Likelihood, which for Gaussian distributed data corresponds to a χ^2 function

$$\chi^2 \; = \; \sum_n^N \frac{(f(x_n, a_k) \; - \; d_n)^2}{\sigma_n^2}$$

For many years the standard approach to selecting the most appropriate model f (or number of model parameters K) was to exploit the observation that for a carefully constructed χ^2 variable (one for which the data is uniform independent normal with known accuracy) the χ^2 per degree of freedom is expected to be unity. For N data points the number of independent degrees of freedom is $N - K$. This statement can be written as

$$< \frac{\chi^2_K}{N-K} > \; = \; 1$$

accordingly we can define a selection rule

$$\chi^2_K \; + \; K \; - \; N \; = \; 0 \qquad (1)$$

as this is expected to reduce monotonically with increasing K, we select a model with the lowest order (K) for which this requirement is satisfied within reasonable statistical limits, and we will refer to this as the Degree of Freedom Method (DOF). This method has solid roots in quantitative statistics, though it is generally understood to be only a first order approximation which may need to be tested using Monte-Carlo.

In the last few decades a new approach to solving this problem was been proposed which can be derived either as a correction to a Likelihood bias w.r.t. unseen test data, or as a measure of *information* [1]. For the correct model it is shown that

$$AIC \; = \; \chi^2_K \; + \; 2K \qquad (2)$$

should be a minimum.

If these two methods are effectively attempting to solve the same problem, the question arises as to how the two methods are related, and in particular are they effectively equivalent.

Analysis

Unfortunately, the two methods can be seen to be mathematically distinct, one (DOF) being sensitive to absolute value, while the other (AIC) to the relative values of χ^2_K. We therefore cannot perform a comparison without having some idea of the behaviour of the χ_K values. To do this we will assume an interpolative function $\chi^2(k)$ Graphically, we can interpret the DOF as the intersection of $\chi^2(k)$ and the line $N - K$, while the AIC method can be interpreted as the value on the curve $AIC(k) \; = \; \chi^2(k) \; + \; 2k$ which has $\frac{\partial AIC(k)}{\partial k} \; = 0$.

We can state a few basic properties of $\chi^2(k)$, the early part of the curve must have values greater than N, and $\chi^2(N) \; = \; 0$. Somewhere between $k = 0$ and $k = N$ the curve must pass through the line $N - k$. The simplest of such functions is piecewise linear (figure 1(a)), and the value of k selected by DOF is immediately defined by

the point of intersection of the first line section and $N - k$. The solutions for AIC will depend upon the specific gradient of the first linear section. If its downwards slope is less than $-2k$ then the optimal solution n will be for $k = 0$, if greater it will be the link point for the two linear sections. At first sight this looks different to the DOF method. However, what must be remembered is that for a meaningful χ^2 once the value of k is greater than or equal to the required number of degrees of freedom we would expect $\chi^2/(N - k)$ to be unity, i.e. it will simply follow the $N - k$ line. The details of shape of the other linear section becomes irrelevant at this point in the analysis. Provided that it is monotonically decreasing with a high slope, the point of intersection should actually be on the $N - k$ line in which case both the DOF and AIC results are identical.

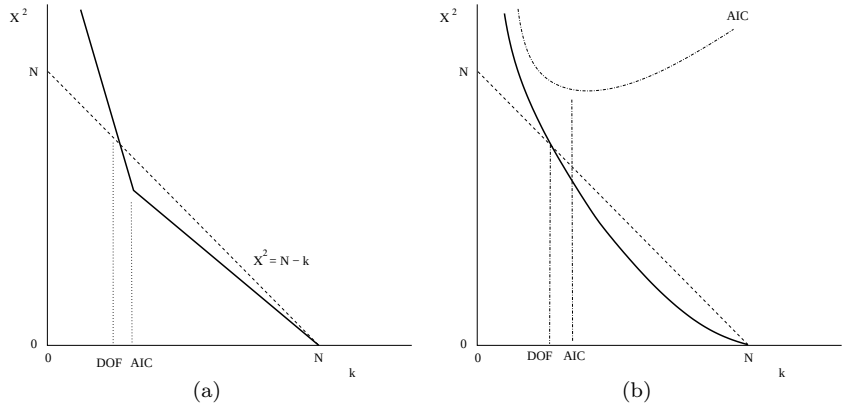

Figure 1: (a) Bi-linear and (b) quadratic approximations to the χ^2 parameter dependency.

If model selection were as simple as the above analysis suggests there would be no problem with determining the best value of k. However, the specific values of χ^2, even under idealised circumstances, do not generate these results. There are several reasons for this; there may be no true model, or even if there is, those fits with too few parameters are adjusted so that the observed values of χ^2 are smaller than we may have expected for the correct parameter values. Equally for models with too many parameters, the values close to the correct model have a larger change in χ^2 than expected, and close to $k = N$ the incremental reduction is less than expected (given a choice of parameters we will always choose the one which accounts for the most variance first, so $\frac{\partial \chi^2}{\partial k}|_{k=N} \rightarrow 0$). As a function, the χ^2 curve will lie below our idealised linear model and may smooth out the kink which indicates an obvious solution.

The simplest function which follows this behaviour is $\chi^2(k) = \alpha(N - k)^2$ with $\alpha > 1/N$ (figure 1(b)). For the DOF method (eq 1), the points of intersection of this function and the line $N - k$ are the solutions of

$$\alpha k^2 - 2\alpha kN + \alpha N^2 + k - N = 0$$

i.e.

$$k = \frac{-(1 - 2\alpha N) \pm \sqrt{(1 - 2\alpha N)^2 - 4\alpha^2(\alpha N^2 - N)}}{2\alpha}$$

which simplifies to

$$k = \frac{-(1 - 2\alpha N) \pm \sqrt{1}}{2\alpha}$$

giving values of $k = N - 1/\alpha$ and $k = N$.

We can determine the solution for the AIC (eq 2) method via differentiation, for which

$$2\alpha(N - k) - 2 = 0 \quad \rightarrow k = N - 1/\alpha$$

Therefore, for a purely quadratic behaviour of $\chi^2(k)$ the AIC and DOF methods are once again identical.

The above analysis is for two highly specific theoretical forms, what we now need to remember is that χ^2 estimates are noisy. This allows us to say that if a particular set of values can be seen to approximate either of the two above models, then we have no reason to expect that the AIC or DOF methods will be capable of telling us anything significantly different with regard to model selection.

Conclusions

Clearly, real χ^2_K behaviours are (in detail) more complex than either of these two models, but this analysis seems to indicate that **to a large extent DOF and AIC are equivalent**. In particular an AIC solution should be consistent wth a DOF test. At some level however we would start to see differences between the results. Then we would need to decide which of the approaches we would be most willing to accept.

At this point we need to go back to underlying definitions to select our preferred method. Personally, I would be swayed by arguments based upon optimal prediction, but less influenced by analogies to information and entropy (which have no real legitimacy with regard to parameter estimation). However, the AIC approach is also consistent with an estimated value of χ^2 which would be generated by N indepenant data samples, i.e. generalisation. Regardless of its origins, the AIC approach seems better to embody this ideal, as for DOF method all models with more than the minimum number of parameters are equally good, it contains no intrinsic behaviour of parsimony unless it is separately enforced by the user (i.e. choose the lowest order satisfactory result).

By the very nature of the approach (computing a zero crossing) the DOF method is likely to be unstable in the presence of noise on computed χ^2_k values, while the AIC solution is more clearly defined. An invalid estimate of the noise may prevent the DOF from generating a solution for any $k < N$. The AIC approach will always indicate a solution even in these circumstances, albeit incorrect. For very large N and small k we may find that AIC has just as much dificulty finding a definite model order as DOF. As both tests are simple, it would be sensible to use AIC and also confirm that the chosen k generates the required χ^2/DOF, or to use the latter to rescale the assumed error model when there is a problem.

Some Observations

It is wrong to interpret an extrapolated function of discrete χ^2_k values as anything too meaningful, and we have taken this approach here as a means to perform an analysis.

However,

if it were really the case that we believed that the DOF and AIC had to be consistent, and

if it were to prove the case that the quadratic interpolation function is the most general polynomial model for which both solutions agree, and

if we were to find that that estimated χ^2 values (which are necessarily stochastic) are well interpolated by a quadratic;

then a case could then be made for fitting the χ^2 dependency with a quadratic and taking the solutions for K around $k = N - 1/\alpha$.

In these cases fitting the interpolating χ^2_K function might then be expected to give more reliable models than the solution based on individual values.

References

[1] Akaike H., A New Look at the Statistical Model Identification, *IEEE Trans. Automatic Control*, 19(6):716-723, 1974.