ADAPTING AIC TO CONDITIONAL MODEL SELECTION

Thijs van Ommen

Centrum Wiskunde & Informatica (CWI),
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands, Thijs.van.Ommen@cwi.nl

ABSTRACT

In statistical settings such as regression and time series, we can condition on observed information when predicting the data of interest. For example, a regression model explains the dependent variables $y_1, \ldots, y_n$ in terms of the independent variables $x_1, \ldots, x_n$. When we ask such a model to predict the value of y_{n+1} corresponding to some given value of x_{n+1}, that prediction's accuracy will vary with x_{n+1}. Existing methods for model selection do not take this variability into account, which often causes them to select inferior models.

One widely used method for model selection is AIC (Akaike's Information Criterion [1]), which is based on estimates of the KL divergence from the true distribution to each model. We propose an adaptation of AIC that takes the observed information into account when estimating the KL divergence, thereby getting rid of a bias in AIC's estimate.

1. A BIAS IN AIC

The principle underlying AIC and many subsequent criteria is that model selection methods should find the model g which minimizes

$$-2\,\mathrm{E}_{\mathbf{U}}\,\mathrm{E}_{\mathbf{V}} \log g(\mathbf{V} \mid \hat{\theta}(\mathbf{U})), \tag{1}$$

where $\hat{\theta}$ represents the maximum likelihood estimator in that model, and both random variables are independent samples of n data points each, both following the true distribution of the data. The inner expectation is the KL divergence from the true distribution to $g(\cdot \mid \hat{\theta}(\mathbf{U}))$ up to a constant which is the same for all models. The quantity (1) can be seen as representing that we first estimate the model's parameters using a random sample $\mathbf{U}$, then judge the quality of this estimate by looking at its performance on an independent, identically distributed sample $\mathbf{V}$.

In regression, time series, and other settings, the data points consist of two parts $u_i = (x_i, y_i)$, and the models are sets of distributions on the *dependent variable* $\mathbf{y}$ conditioned on the *independent variable* x (which may or may not be random). We call these *conditional* models. Then (1) can be adapted in two ways: as the extra-sample error

$$-2\,\mathrm{E}_{\mathbf{Y}|X}\,\mathrm{E}_{\mathbf{Y}'|X'} \log g(\mathbf{Y}' \mid X', \hat{\theta}(X, \mathbf{Y})), \tag{2}$$

and, replacing both X and X' by a single variable X, as the in-sample error

$$-2\,\mathrm{E}_{\mathbf{Y}|X}\,\mathrm{E}_{\mathbf{Y}'|X} \log g(\mathbf{Y}' \mid X, \hat{\theta}(X, \mathbf{Y})). \tag{3}$$

The standard expression behind AIC (1) makes no reference to X or X', which leads a straightforward derivation of AIC for a conditional model to make the tacit assumption $X = X'$, so that standard AIC estimates the in-sample error. This applies for instance to the well-known form of AIC for linear models, i.e. the residual sum of squares with a penalty of $2k$, where k is the model's order.

However, the extra-sample error (2) is more appropriate as a measure of the expected performance on new data. Using the in-sample error (3) instead results in a biased estimate of this performance. As the bias gets worse for larger models, this will lead to inferior model selection.

2. AN UNBIASED ADAPTATION

To get an estimator for (2), we do not make any assumptions about the process generating X and X' (it may not even be random) but treat their values as given. We denote the number of data points in X and X' by n and n', respectively. In the case of simple linear regression with fixed variance, a derivation similar to AIC's leads to a penalty term of $k + \kappa_{X'}$ in place of AIC's $2k$, where

$$\kappa_{X'} = \frac{n}{n'} \operatorname{tr}\left[{X'}^{\top} X' (X^{\top} X)^{-1}\right],$$

where X and X' represent design matrices. Similarly, a small sample corrected version analogous to AICc [2] can be derived and has penalty

$$k + \kappa_{X'} + \frac{(k + \kappa_{X'})(k+1)}{n - k - 1}.$$

3. FOCUSED AIC FOR PREDICTION

If our goal is prediction, then the value X used in our derivation corresponds to the data we have observed already, and X' may be replaced by the single point x for which we need to predict the corresponding $\mathbf{y}$. This justifies treating X and X' as given in this practical setting. Thus we use x already at the stage of model selection, whereas standard methods for model selection only use it after selecting a model, to find the distribution of $\mathbf{y}$ conditioned on that x. Then for the linear model with fixed

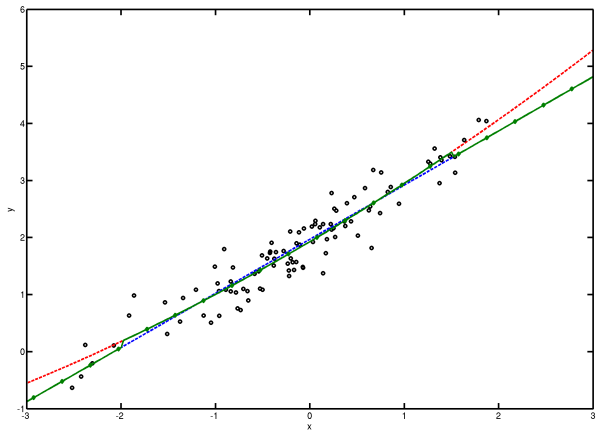

Figure 1. Example illustrating the result of applying FAIC to a sample of 100 data points. There are three models: the constant, linear, and quadratic functions; the true distribution uses a linear function. The choice of FAIC is marked in green: it selects a quadratic (red) function for x close to many observed data points, and a linear (blue) function elsewhere.

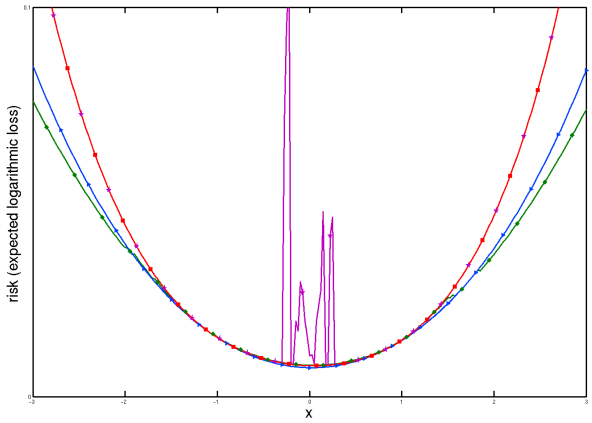

Figure 2. Average performance of different model selection methods as a function of x. Our FAIC (in green) outperforms the other methods for extreme x and is competetive otherwise; AIC (red) overfits especially for extreme x; BIC (Bayesian Information Criterion, blue) is less likely to overfit than AIC; FIC (Focused Information Criterion, purple) is similar to AIC but selects a constant function in the center.

variance, κ_x becomes

$$\kappa_x = \frac{n}{n'} \operatorname{tr}[xx^\top (X^\top X)^{-1}] = nx^\top (X^\top X)^{-1} x;$$

for unknown variance it becomes this value plus one.

We name this method Focused AIC. The term "focus" was first used by Claeskens and Hjort's [3] to describe a model selection method that focuses on a parameter of interest when selecting a model. The behaviour of FAIC is illustrated in Figure 1.

4. EXPERIMENTAL RESULTS

Simulation experiments with linear regression models indicate that our method outperforms AIC in terms of logarithmic (or squared) loss in many situations. Representative results are shown in Figure 2.

5. REFERENCES

[1] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd International Symposium on Information Theory*, Tsahkadsor, Armenian SSR, Sep. 1971, pp. 267–281.

[2] C. M. Hurvich and C-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, pp. 297–307, Jun. 1989.

[3] G. Claeskens and N. L. Hjort, "The focused information criterion," *Journal of the American Statistical Association*, vol. 98, pp. 900–916, Dec. 2003.