KULLBACK-LEIBLER INFORMATION THEORY

A BASIS FOR MODEL SELECTION AND INFERENCE

Kullback-Leibler Information or Distance

$$I(f, g) = \int f(x) \log \left(\frac{f(x)}{g(x|\theta)} \right) dx,$$

$I(f, g)$ is the "information" lost when model g is used to approximate full truth f.

$I(f, g)$ is the "distance" between a model and full truth.

<u>Glimpse into derivation from K-L information to AIC:</u>

$$I(f, g) = \int f(x) \log (f(x))\, dx \;-\; \int f(x) \log (g(x \mid \theta)) dx.$$

$$I(f, g) = \mathrm{E}_f [\log(f(x))] - E_f [\log(g(x \mid \theta))].$$

$$I(f, g) = \text{Constant} - \mathrm{E}_f [\log(g(x \mid \theta))],$$

or

$$I(f, g) - \text{Constant} = - \mathrm{E}_f [\log(g(x \mid \theta))].$$

The term $\mathrm{E}_f[\log(g(x \mid \theta))]$ becomes the quantity of interest, but cannot be estimated. Akaike found that its expectation

$$\mathrm{E}_f\mathrm{E}_f[\log(g(x \mid \theta))]$$

can be estimated! An asymptotically unbiased estimator of the relative, expected K-L information is

$$\log(\mathcal{L}(\hat{\theta} \mid \text{data})) - K,$$

where K is the number of estimable parameters in the model.

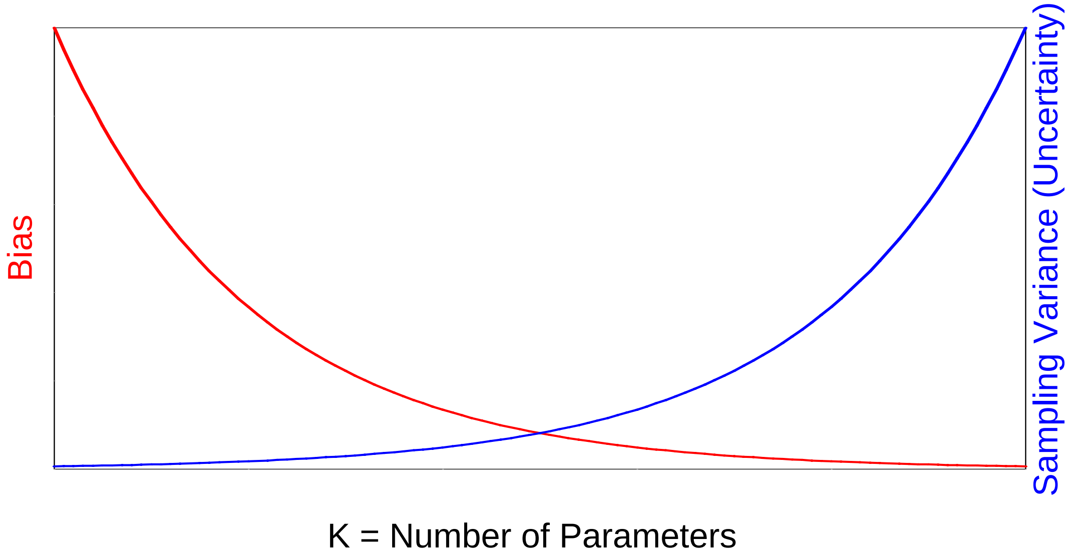
Akaike (1973) then defined "*an information criterion*" (AIC) by multiplying by -2 ("taking historical reasons into account") to get

$$\text{AIC} = -2\log(\mathcal{L}(\hat{\theta} \mid \text{data})) + 2K.$$

Thus, one should select the model that yields the smallest value of AIC because this model is estimated to be "closest" to the unknown reality that generated the sample data, from among the candidate models considered. There are R models in the set: g_1, g_2, ..., g_R.

This seems a very natural, simple concept; select the fitted approximating model that is estimated, on average, to be closest to the unknown truth, f.

If all the models in the set are poor, AIC attempts to select the best approximating model of those in the candidate set and ranks the rest. Statistics such as R^2 are useful here. Thus, every effort must be made to assure that the set of models is well founded. Much more hard thinking is called for here.


Bias
Sampling Variance (Uncertainty)
K = Number of Parameters

AIC Differences

AIC contains several types of constants and is a function of sample size, we recommend computing (and presenting in publications) the AIC differences (in addition to the actual AIC values),

$$\Delta_i = \text{AIC}_i - \text{minAIC},$$

where minAIC is the smallest AIC value in the set. Thus, that best model has $\Delta_{\min} = 0$. The larger Δ_i is, the less plausible is the fitted model $g_i(x \mid \hat{\theta})$ as being the K-L best model for samples such as the data one has. The simple differencing leading to Δ_i can be used with AICc and QAICc, as explained below.

Important Refinements to AIC

A Second Order AIC

Akaike derived an asymptotically unbiased estimator of K-L information, however, AIC may perform poorly if there are too many parameters in relation to the size of the sample.

A small-sample (second order) bias adjustment which led to a criterion that is called AIC$_c$ (Sugiura (1978) and Hurvich and Tsai (1989)),

$$\text{AIC}_c = -2\log(\mathcal{L}(\hat{\theta})) + 2K\left(\frac{n}{n-K-1}\right),$$

This can be rewritten equivalently as

$$\text{AIC}_c = -2\log(\mathcal{L}(\hat{\theta})) + 2K + \frac{2K(K+1)}{n-K-1},$$

or, equivalently,

$$\text{AIC}_c = \text{AIC} + \frac{2K(K+1)}{n-K-1},$$

where *n* is sample size. Because AIC and AIC$_c$ converge when sample size is large, <u>one should always use AICc.</u>

Modification to AIC for Overdispersed Count Data

Count data have been known not to conform to simple variance assumptions based on binomial or multinomial distributions. If the sampling variance exceeds the theoretical (model based) variance, the situation is called "overdispersion" or "extra-binomial variation."

The first useful approximation is based on a single variance inflation factor (c) which can be estimated from the goodness-of-fit chi-square statistic (χ^2) of the global model and its degrees of freedom,

$$\hat{c} = \chi^2/df.$$

Given overdispersed data and a variance inflation factor, $\hat{c}$, three things should be done/considered:

1. **empirical estimates of sampling variances ($var_e(\hat{\theta}_i)$) and covariances ($cov_e(\hat{\theta}_i, \hat{\theta}_j)$) should be inflated by multiplying by the estimate of c. Note, the standard errors are inflated by the square root of $\hat{c}$.**

2. **model selection criteria need slight modification,**

$$\text{QAIC}_c = -\left[2\log(\mathcal{L}(\hat{\theta}))/\hat{c}\right] + 2K + \frac{2K(K+1)}{n-K-1},$$

$$= \text{QAIC} + \frac{2K(K+1)}{n-K-1}.$$

3. **the parameter count, K must be increased by 1 (for the estimation of $\hat{c}$).**

Some History

Akaike (1973) considered AIC and its information theoretic foundations "... a natural extension of the classical maximum likelihood principle." Interestingly, Fisher (1936) anticipated such an advance over 60 years ago when he wrote,

> **"... an even wider type of inductive argument may some day be developed, which shall discuss methods of assigning from the data the functional form of the population."**

Science Hypotheses and Modeling

A well thought out global model (where applicable) is important and substantial prior knowledge is required during the entire survey or experiment, including the clear statement of the question to be addressed and the collection of the data. This prior knowledge is then carefully input into the development of the set of candidate models.

Without this background science, the entire investigation should probably be considered only very preliminary.

MULTIMODEL INFERENCE

Making Inferences From More Than a Single Model

The Likelihood of a Model

We can extend the concept of the likelihood of the parameters given a model and data,

$$\mathcal{L}(\underline{\theta} \mid \underline{x}, g_i),$$

to a concept of the likelihood of the model given the data ($\underline{x}$),

$$\mathcal{L}(g_i \mid \underline{x}) \propto exp(-\tfrac{1}{2}\Delta_i).$$

[Note the $-\frac{1}{2}$ here just erases the fact that Akaike multiplied through by –2 to define his AIC]

To better interpret these relative likelihoods of models given the data and the set of *R* models, we normalize them to be a set of positive "Akaike weights" adding to 1:

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^{R}\exp(-\frac{1}{2}\Delta_r)}.$$

Model Probabilities

A given w_i is considered as the weight of evidence in favor of model *i* as being the actual K-L best model in the set.

These are termed model probabilities. In fact, they are also formally Bayesian posterior model probabilities (Burnham and Anderson 2004). So, w_i is the probability that model *i* is *the* actual K-L best model in the set.

The bigger the Δ_i value, the smaller the weight.

The bigger a Δ_i is, the less plausible is model *i* as being the actual K-L best model of full reality, based on the design and sample size used.

Evidence Ratios

Evidence for pairs of hypotheses or models can be judged via an *evidence ratio*. Such ratios are invariant to other models in or out of the model set. Evidence ratios between model *i* and model *j* are trivial to compute, and can be gotten as,

$$\mathcal{L}(g_i \mid \underset{\sim}{x}) / \mathcal{L}(g_j \mid \underset{\sim}{x})$$

or

$$w_i/w_j \ .$$

Most often, one wants the evidence ratio between the best model (*b*) and the *j*th model, w_b/w_j . There is a striking nonlinearity between the evidence ratio and the Δ_i values,

Δ_i	Evidence ratio
2	2.7
4	7.4
8	54.6
10	148.4
15	1,808.0
20	22,026.5
50	72 billion.

Model Averaging

Sometimes there are several models that seem plausible, based on the AIC$_c$ or QAIC$_c$ values. In this case, there is a formal way to base inference on more than a single model. A model averaged estimator is

$$\hat{\bar{\theta}} = \sum_{i=1}^{R} \hat{w}_i \, \hat{\theta}_i \, ,$$

where *i* indexes the models.

Note, model averaging is needed to compute the unconditional variance (below).

"Unconditional" Estimates of Precision

The precision of an estimator should ideally have 2 variance components:

(1) the conditional sampling variance, given a model $\left(\hat{\text{v}}\text{ar}(\hat{\theta}_i \mid g_i)\right)$, and

(2) variation associated with model selection uncertainty. Thus,

$$\hat{\text{v}}\text{ar}(\hat{\bar{\theta}}) =$$

$$\sum_{i=1}^{R} \hat{w}_i \left\{ \hat{\text{v}}\text{ar}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \hat{\bar{\theta}})^2 \right\},$$

where,

$$\hat{\bar{\theta}} = \sum_{i=1}^{R} \hat{w}_i \hat{\theta}_i$$

The estimated conditional standard error is,

$$\hat{\text{s}}\text{e}(\hat{\bar{\theta}}) = \sqrt{\hat{\text{v}}\text{ar}(\hat{\bar{\theta}})}.$$

Unconditional Confidence Intervals

A simple approximation to a $(1-\alpha)100\%$ unconditional confidence interval is just,

$$\hat{\theta}_i \pm z_{1\text{-}\alpha/2}\,\widehat{\text{se}}(\hat{\theta}_i),$$

where *i* is over the *R* models.

Then,

$$\widehat{\text{se}}(\hat{\theta}_i) = \sqrt{\widehat{\text{var}}(\hat{\theta}_i)}\,.$$

Such unconditional confidence intervals can be set around a single $\hat{\theta}$ or a model averaged estimate $\hat{\theta}$.

The word "unconditional" is perhaps unfortunate as the estimates of precision are still conditioned on the set of models. They are "unconditional" in the sense that they are not conditioned on a single (usually best) model.

Summary

The Principle of Parsimony provides a *conceptual guide* to model selection.

Expected K-L information provides an *objective criterion*, based on a deep theoretical justification.

AIC$_c$ and QAIC$_c$ provide a *practical method* for model selection and associated data analysis and are estimates of expected, relative K-L information.

AIC, AIC$_c$ and QAIC represent an extensions of classical likelihood theory, are applicable across a very wide range of scientific questions, and are quite simple to compute and interpret in practice.