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 g is used to approximate f.
$I(f, g)$ is the "distance" between a model and truth (i.e., full reality).

$I(f, g)$ can be written equivalently as

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

Note, each of the two terms on the right of the above expression is a statistical expectation with respect to f (truth). Thus,

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

The first expectation $\mathrm{E}_f[\log(f(x))]$ is a constant across models, thus,

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

or

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

The term $\left(I(f, g) - \text{Constant}\right)$ is a *relative*, directed distance between f and g; now if one could compute or estimate $\mathbf{E}_f[\log(g(x \mid \theta))]$. Thus, $\mathbf{E}_f[\log(g(x \mid \theta))]$ 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, g. Akaike's finding of a relation between the relative K-L distance and the maximized log-likelihood has allowed major

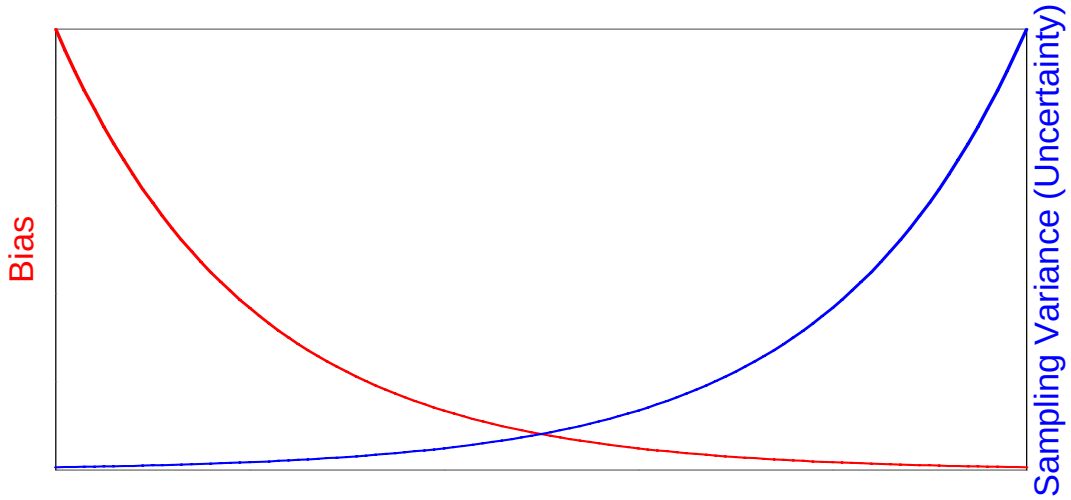
practical and theoretical advances in model selection and the analysis of complex data sets (see Stone 1982, Bozdogan 1987, and deLeeuw 1992).

Akaike (1973) then defined "*an information criterion*" (AIC) by multiplying by -2,

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

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 data, from among the candidate models considered. The model set is defined as $g_1, g_2, \ldots, g_R$. Thus, there are R models in the candidate set.

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 to help quantify that some models are, at least, of some use. Thus, every effort must be made to assure that the set of models is well founded.



AIC Differences

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

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

where minAIC is the smallest AIC value in the set. Thus, that best model has $\Delta_{\text{min}} = 0$. 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)),

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

where the penalty term is multiplied by the correction factor $n/(n–K–1)$. This can be rewritten equivalently as

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

or, equivalently,

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

where n is sample size. Because AIC and AIC_c converge when sample size is large, one can **always use AICc.**

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." Cox and Snell (1989) discuss modeling of count data and note that 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,

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

The variance inflation factor should be estimated from the global model. There are most effective, computer intensive approaches to the estimation of the variance inflation factor (not ocvered in this introduction).

Given overdispersed data and an estimated variance inflation factor, $\hat{c}$, three things must 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 inflacted by multiplying by the estimate of c. Note, the standard error must be inflated by the square root of $\hat{c}$.

2. modified model selection criteria must be used,

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

and

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

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

3. the parameter count, K must be increased by 1 (for the estimation of 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.*

Making Inferences From More Than a Single 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:

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

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. Again, most applications of the above will be using AICc or QAICc.

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,

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

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 $\left(\hat{\text{var}}(\hat{\theta}_i \mid g_i) \right)$, and
(2) variation associated with model selection uncertainty. Thus,

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

where,

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

Obviously, the estimated conditional standard error is,

$$\hat{\text{se}}(\hat{\bar{\theta}}) = \sqrt{\hat{\text{var}}(\hat{\bar{\theta}})} .$$

Unconditional Confidence Intervals

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

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

where *i* is over the *R* models.
Of course,

$$\hat{\mathrm{se}}(\bar{\theta}_i) = \sqrt{\hat{\mathrm{var}}(\bar{\theta}_i)}.$$

Such unconditional confidence intervals can be set around a single $\hat{\theta}$ or a model averaged estimate $\bar{\theta}$. When there is no model selection then an interval, conditional on model *i* is the usual,

$$\hat{\theta}_i \pm t_{df,1-\alpha/2}\, \hat{\mathrm{se}}(\hat{\theta}_i \mid g_i),$$

where it is clear what the degrees of freedom (*df*) are for the *t*-distribution.

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 QAICc 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.

Some References On Model Selection

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. Pages 267-281 *in* B. N. Petrov, and F. Csaki, (Eds.) *Second International Symposium on Information Theory*. Akademiai Kiado, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control AC* **19**, 716-723.

Akaike, H. (1977). On entropy maximization principle. Pages 27-41 *in* P. R. Krishnaiah (Ed.) *Applications of statistics*. North-Holland, Amsterdam.

Akaike, H. (1981a). Likelihood of a model and information criteria. *Journal of Econometrics* **16**, 3-14.

Akaike, H. (1981b). Modern development of statistical methods. Pages 169-184 *in* P. Eykhoff (Ed.) *Trends and progress in system identification.* Pergamon Press, Paris.

Akaike, H. (1983a). Statistical inference and measurement of entropy. Pages 165-189 *in* G. E. P. Box, T. Leonard, and C-F. Wu (Eds.) *Scientific inference, data analysis, and robustness*. Academic Press, London.

Akaike, H. (1983b). Information measures and model selection. *International Statistical Institute* **44**, 277-291.

Akaike, H. (1985). Prediction and entropy. Pages 1-24 *in* A. C. Atkinson, and S. E. Fienberg (Eds.) *A celebration of statistics.* Springer, New York, NY.

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. Pages 610-624. *in* S. Kotz, and N. L. Johnson (Eds.) *Breakthroughs in statistics,* Vol. 1. Springer-Verlag, London.

Akaike, H. (1994). Implications of the informational point of view on the development of statistical science. Pages 27-38 *in* H. Bozdogan, (Ed.) *Engineering and Scientific Applications*. Vol. 3, Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach. Kluwer Academic Publishers, Dordrecht, Netherlands.

Anonymous. (1997). *The Kullback Memorial Research Conference.* Statistics Department, The George Washington University, Washington, D. C. 36pp.

Anderson, D. R., and Burnham, K. P. (1999). General strategies for the collection and analysis of ringing data. *Bird Study* 46 (Suppl.) S261-270.

Anderson, D. R., and Burnham, K. P. (1999). Understanding information criteria for selection among capture-recapture or ring recovery models. *Bird Study* 46 (Suppl.) S14-21.

Anderson, D. R., and K. P. Burnham. 2002. Avoiding pitfalls when using information-theoretic methods. *Journal of Wildlife Management* 66:912-918.

Anderson, D. R., Burnham, K. P., and White, G. C. (1994). AIC model selection in overdispersed capture-recapture data. *Ecology* **75**, 1780-1793.

Anderson, D. R., Burnham, K. P., and White, G. C. (1998). Comparison of AIC and CAIC for model selection and statistical inference from capture-recapture studies. *Journal of Applied Statistics* **25**, 263-282.

Anderson, D. R., K. P. Gould, W. R. Gould, and S. Cherry. 2001. Concerns about finding effects that are actually spurious. *The Wildlife Society Bulletin* 29:311-316.

Anderson, D. R., W. A. Link, D. H. Johnson, and K. P. Burnham. 2001. Suggestions for presenting results of data analysis. *Journal of Wildlife Management* 65:373-378.

Azzalini, A. (1996). *Statistical inference – based on the likelihood.* Chapman and Hall, London.

Boltzmann, L. (1877). Uber die Beziehung zwischen dem Hauptsatze der mechanischen Warmetheorie und der Wahrscheinlicjkeitsrechnung respective den Satzen uber das Warmegleichgewicht. *Wiener Berichte* **76**, 373-435.

Box, J. F. (1978). *R. A. Fisher: the life of a scientist.* John Wiley and Sons, New York, NY. 511pp.

Burnham, K. P., and D. R. Anderson. 2001. Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research* 28:111-119.

Burnham, K. P., and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information theoretic approach*. 2nd Ed., Springer-Verlag, New York. 488pp.

Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Methods in Sociological Research.* 33:261-304.

Burnham, K. P., Anderson, D. R., and White, G. C. (1994). Evaluation of the Kullback–Leibler discrepancy for model selection in open population capture-recapture models. *Biometrical Journal* **36**, 299-315.

Burnham, K. P., White, G. C., and Anderson, D. R. (1995a). Model selection in the analysis of capture-recapture data. *Biometrics* **51**, 888-898.

Burnham, K. P., Anderson, D. R., and White, G. C. (1995b). Selection among open population capture-recapture models when capture probabilities are heterogeneous. *Journal of Applied Statistics* **22**, 611-624.

Burnham, K. P., Anderson, D. R., and White, G. C. (1996). Meta-analysis of vital rates of the Northern Spotted Owl. *Studies in Avian Biology* **17**, 92-101.

Chamberlain, T. C. (1890). The method of multiple working hypotheses. *Science* **15**, 93.

Chatfield, C. (1991). Avoiding statistical pitfalls (with discussion). *Statistical Science* **6**, 240-268.

Chatfield, C. (1995a). *Problem solving: a statistician's guide.* Second edition. Chapman and Hall, London. 325pp.

Chatfield, C. (1995b). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society,* Series A **158**, 419-466.

Cover, T. M., and Thomas, J. A. (1991). *Elements of information theory.* John Wiley and Sons, New York, NY. 542pp.

de Leeuw, J. (1988). Model selection in multinomial experiments. Pages 118-138 *in* T. K. Dijkstra (Ed.) *On model uncertainty and its statistical implications.* Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, New York, NY.

de Leeuw, J. (1992). Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. Pages 599-609 *in* S. Kotz, and N. L. Johnson (Eds.) *Breakthroughs in statistics.* Vol. 1. Springer-Verlag, London.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Royal Society of London. *Philosophical Transactions* (Series A) **222**, 309-368.

Hurvich, C. M., and Tsai, C-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76,** 297-307.

Kullback, S. (1959). *Information theory and statistics.* John Wiley and Sons, New York, NY.

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79-86.

Lebreton, J-D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monograph* **62,** 67-118.

Parzen, E. (1994). Hirotugu Akaike, statistical scientist. Pages 25-32 *in* H. Bozdogan (Ed.) *Engineering and Scientific Applications.* Vol. 1,

Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach. Kluwer Academic Publishers, Dordrecht, Netherlands.

Parzen, E., Tanabe, K, and Kitagawa, G. (Eds.) (1998). *Selected papers of Hirotugu Akaike*. Springer-Verlag Inc., New York, NY.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). *Akaike information criterion statistics.* KTK Scientific Publishers, Tokyo.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**, 379-423 and 623-656.

Stone, C. J. (1982). Local asymptotic admissibility of a generalization of Akaike's model selection rule. *Annals of the Institute of Statistical Mathematics* Part A **34**, 123-133.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods*. **A7**, 13-26.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku* (Mathematic Sciences) **153**, 12-18. (In Japanese).

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439-447.