INFORMATION THEORY AND LOG-LIKELIHOOD MODELS: A BASIS FOR MODEL SELECTION AND INFERENCE

Full reality cannot be included in a model; thus we seek a good model to *approximate* the effects or factors supported by the empirical data. The selection of an appropriate approximating model is critical to statistical inference from many types of empirical data. Information theory (see Guiasu 1977 and Cover and Thomas 1991) and, in particular the Kullback-Leibler (Kullback and Liebler 1951) “distance," or “information" forms the deep theoretical basis for data-based model selection. Akaike (1973) found a simple relationship between expected Kullback-Leibler information and Fisher's maximized log-likelihood function (see deLeeuw 1992 for a brief review). This relationship leads to a simple, effective, and very general methodology for selecting a parsimonious model for the analysis of empirical data.

Some Background

Well over a century ago measures were derived for assessing the “distance" between two models or probability distributions. Most relevant here is Boltzmann's (1877) concept of generalized entropy in physics and thermodynamics (see Akaike 1985 for a brief review). Shannon (1948) employed entropy in his famous treatise on communication theory. Kullback and Leibler (1951) derived an information measure that happened to be the negative of Boltzmann's entropy: now referred to as the Kullback-Leibler (K-L) distance. The motivation for Kullback and Leibler's work was to provide a rigorous definition of “information" in relation to Fisher's “sufficient statistics." The K-L distance has also been called the K-L discrepancy, divergence, information and number – these terms are synonyms, we tend to use *distance* or *information* in the material to follow.

The Kullback-Leibler distance can be conceptualized as a directed “distance" between two models, say f and g (Kullback 1959). Strictly speaking this is a measure of “discrepancy"; it is not a simple distance because the measure from f to g is not the same as the measure from g to f – it is a directed or oriented distance. The K-L distance is perhaps the most fundamental of all information measures in the sense of being derived from minimal assumptions and its additivity property. The K-L distance between models is a *fundamental quantity* in science and information theory (see Akaike 1983) and is the logical basis for model selection as defined by Akaike. In the heuristics given here, we will assume the models in question are continuous probability distributions denoted as f and g. Biologists are familiar with the normal, log-normal, gamma and other continuous distributions. We will, of course, not limit ourselves to these common, simple types.

It is useful to think of f as full reality and let it have (conceptually) an infinite number of parameters. This “crutch" of infinite dimensionality at least keeps the concept of reality even though it is in some unattainable perspective.

Let g be the approximating model being compared to (measured against) f. We use x to denote the data being modeled and θ to denote the parameters in the approximating model g. We use $g(x)$ as an approximating model, whose parameters must be estimated from these data (in fact, we will make this explicit using the notation $g(x \mid \theta)$, read as “the approximating model g for data x given the parameters θ). If the parameters of the model g have been estimated, using ML or LS methods, we will denote this by $g(x \mid \hat{\theta})$. Generally, in any real world problem, the model $g(x \mid \theta)$ is a function of sample data (often multivariate) and the number of parameters (θ) in g might often be of high dimensionality. Finally, we will want to consider a set of approximating models as candidates for the representation of the data; this set of models is denoted $\{g_i(x \mid \theta), i = 1, ..., R\}$. **It is critical that this set of models be defined prior to probing examination of the data (i.e., no data dredging).**

The Kullback-Leibler Distance or Information

The K-L distance between the models *f* and *g* is defined for continuous functions as the (usually multi-dimensional) integral

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

where log denotes the natural logarithm. Kullback and Leibler (1951) developed this quantity from “information theory," thus they used the notation *I*(*f*, *g*);

***I*(*f*, *g*) is the “information" lost when *g* is used to approximate *f*.**

Of course, we seek an approximating model that loses as little information as possible; this is equivalent to minimizing *I*(*f*, *g*), over *g*. *f* is considered to be given (fixed) and only *g* varies over a space of models indexed by θ. An equivalent interpretation of minimizing *I*(*f*, *g*) is finding an approximating model that is the “shortest distance" away from truth. We will use both interpretations as both seem useful.

The expression for the K-L distance in the case of discrete distributions such as the Poisson, binomial or multinomial, is

$$I(f, g) = \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{\pi_i}\right).$$

Here, there are *k* possible outcomes of the underlying random variable; the true probability of the i^{th} outcome is given by p_i, while the $\pi_1, \dots, \pi_k$ constitute the approximating probability distribution (i.e., the approximating model). In the discrete case, we have $0 < p_i < 1$, $0 < \pi_i <1$ and $\sum p_i = \sum \pi_i = 1$. Hence, here *f* and *g* correspond to the *p* and π, respectively.

The material above makes it obvious that both *f* and *g* (and their parameters) must be known to compute the K-L distance between these 2 models. We see that this requirement is diminished as *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, the K-L distance (above) can be expressed as a difference between two expectations,

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

each with respect to the true distribution f. This last expression provides easy insights into the derivation of AIC. The important point is that the K-L distance $I(f, g)$ is a measure of the directed "distance" between the probability models f and g.

The first expectation $\mathrm{E}_f[\log(f(x))]$ is a constant that depends only on the unknown true distribution and it is clearly not known (i.e., we do not know $f(x)$ in actual data analysis). Therefore, treating this unknown term as a constant, only a measure of *relative*, directed distance is possible (Bozdogan 1987, Kapur and Kesavan 1992: 155) in practice. Clearly, if one computed the second expectation, $\mathrm{E}_f[\log(g(x \mid \theta))]$, one could estimate $I(f, g)$, up to a constant (namely $\mathrm{E}_f[\log(f(x))]$),

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

or

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

The term $\Big(I(f, g) - \text{Constant}\Big)$ is a *relative*, directed distance between the two models f and g, if one could compute or estimate $\mathrm{E}_f[\log(g(x \mid \theta))]$. Thus, $\mathrm{E}_f[\log(g(x \mid \theta))]$ becomes the quantity of interest.

In data analysis the model parameters must be estimated and there is usually substantial uncertainty in this estimation. Models based on estimated parameters, hence on $\hat{\theta}$ not θ, represent a major distinction from the case where model parameters would be known. This distinction affects how we must use K-L distance as a basis for model selection. The difference between having θ (we do not) and having $\hat{\theta}$ (we do) is quite important and basically causes us to change our model selection criterion to that of minimizing *expected* estimated K-L distance rather than minimizing known K-L distance (over the set of *Ra priori* models considered).

Thus, we use the *concept* of selecting a model based on minimizing the estimated Kullback-Leibler distance

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

Consequently, we can determine a method to select the model g_i that on average minimizes, over the set of models $g_1, \ldots, g_R$, a very relevant expected K-L distance.

Akaike (1973) showed firstly that the maximized log-likelihood is biased upward as an estimator of the model selection criterion. Second, he showed that under certain conditions (these conditions are important, but quite technical), that this bias is approximately equal to K, the number of estimable parameters in the approximating model. Thus, an approximately unbiased estimator of the relative, expected K-L information is

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

This result is equivalent to

$$\log(\mathcal{L}(\hat{\theta} \mid y)) - K = \text{Constant} - \hat{\mathrm{E}}_{\hat{\theta}}\,[\hat{I}(f, g)]$$

or

$$-\log(\mathcal{L}(\hat{\theta} \mid y)) + K = \text{estimated relative expected K-L distance.}$$

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).

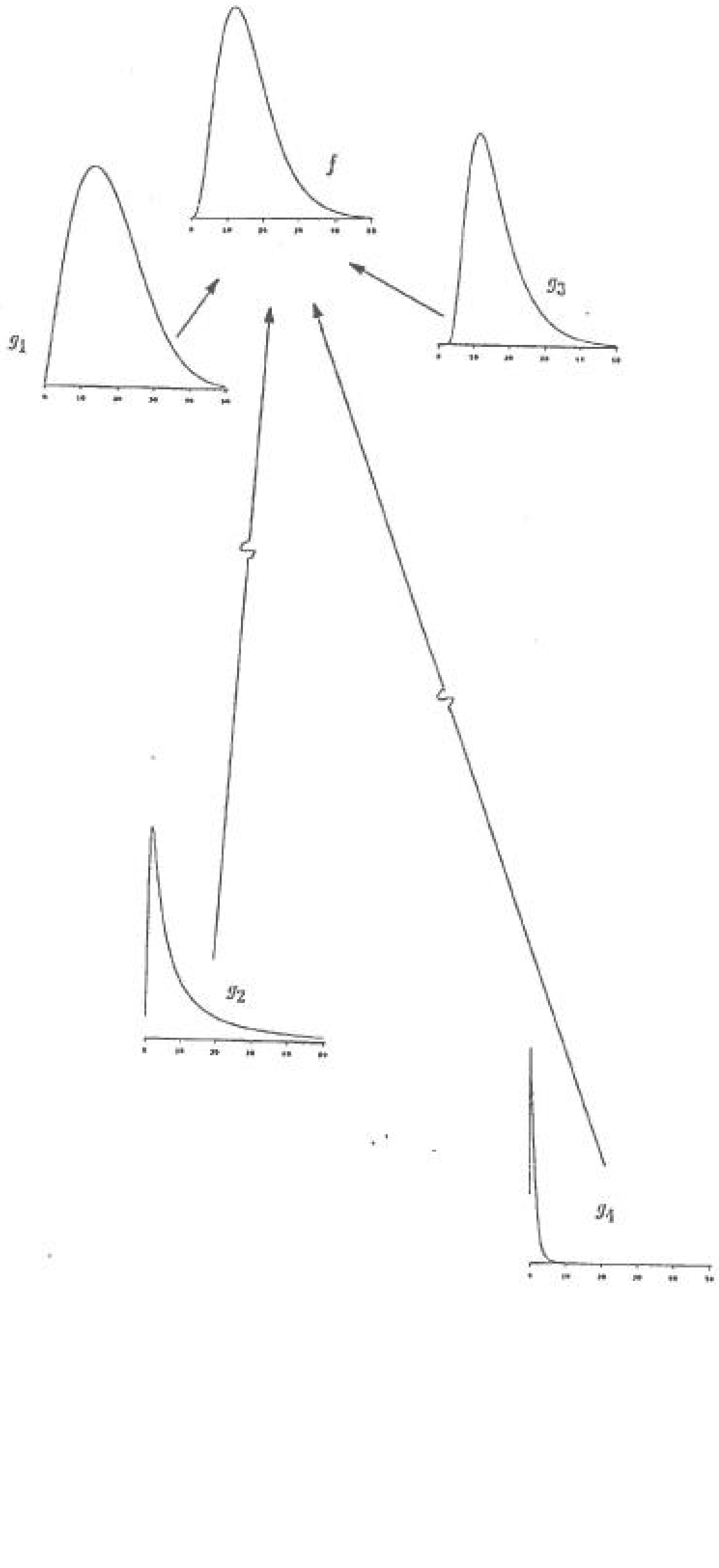
Akaike (1973) then defined "*an information criterion*" (AIC) by multiplying by -2 ("taking historical reasons into account") to get

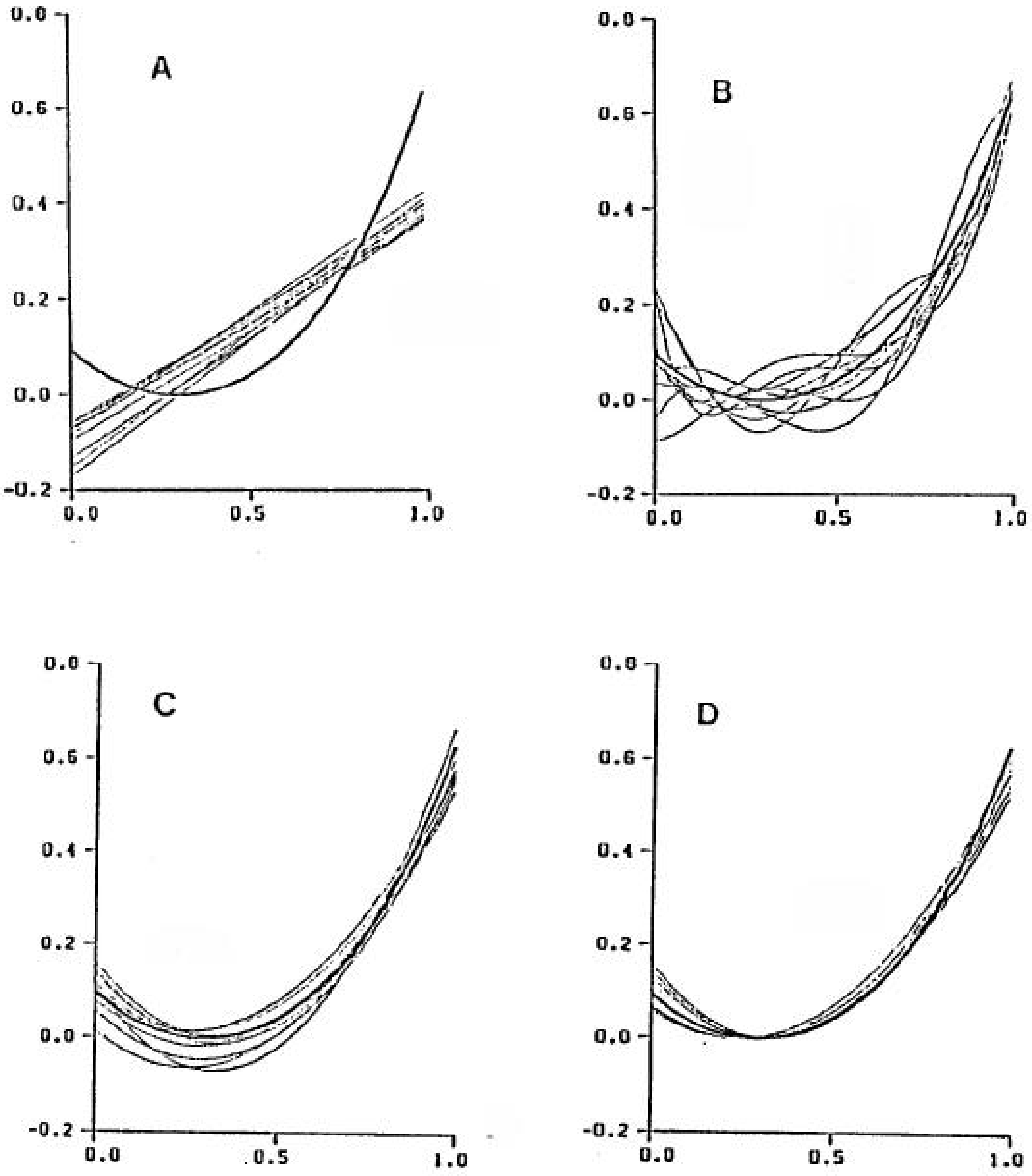
$$\mathbf{AIC} = -2\log(\mathcal{L}(\hat{\theta} \mid y)) + 2K.$$

This has became known as "*Akaike's Information Criterion*" or AIC. Here it is important to note that AIC has a strong theoretical underpinning, based on information theory and Kullback-Leibler information within a realistic data analysis philosophy that no model is true, rather truth as f is far more complex than any model used. Akaike's inferential breakthrough was realizing that a predictive expectation version of the log-likelihood could (as one approach) be used to estimate the relative expected K-L distance between the approximating model and the true generating mechanism. Thus, rather than having a simple measure of the directed distance between two models (i.e., the K-L distance), one has instead an *estimate* of the expected relative, directed distance between the fitted model and the unknown true mechanism (perhaps of infinite dimension) which actually generated the observed data. Because the expectation of the logarithm of $f(x)$ drops out as a constant, independent of the data, AIC is defined without specific reference to a "true model" (Akaike 1985:13).

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

Perhaps none of the models in the set are good, but AIC attempts to select the best approximating model of those in the candidate set. Thus, every effort must be made to assure that the set of models is well founded.


f
g_1
g_3
g_2
g_4


A
0.0
0.6
0.4
0.2
0.0
-0.2
0.0
0.5
1.0
B
0.0
0.6
0.4
0.2
0.0
-0.2
0.0
0.5
1.0
C
0.0
0.6
0.4
0.2
0.0
-0.2
0.0
0.5
1.0
D
0.0
0.6
0.4
0.2
0.0
-0.2
0.0
0.5
1.0

AIC Differences

Because AIC is on an relative scale, we routinely recommend computing (and presenting in publications) the **AIC differences** (rather than the actual AIC values),

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

$$\doteq \text{E}_{\hat{\theta}}[\hat{I}(f, g_i)] - \text{minE}_{\hat{\theta}}[\hat{I}(f, g_i)],$$

over all candidate models in the set. Such differences estimate the relative expected K-L differences between f and $g_i(x \mid \theta)$. These Δ_i values are easy to interpret and allow a quick comparison and ranking of candidate models and are also useful in computing Akaike weights and other quantities of interest.

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. As a rough rule of thumb, models for which $\Delta_i \leq 2$ have substantial support and should receive consideration in making inferences. Models having Δ_i of about 4 to 7 have considerably less support, while models with $\Delta_i > 10$ have either essentially no support, and might be omitted from further consideration, or at least those models fail to explain some substantial explainable variation in the data. If observations are not independent but are assumed to be independent then these simple guidelines cannot be expected to hold.

Important Refinements to AIC

A Second Order AIC

Akaike derived an estimator of the K-L information quantity, however, AIC may perform poorly if there are too many parameters in relation to the size of the sample (Sugiura 1978, Sakamoto et al. 1986). Sugiura (1978) derived a second order variant of AIC that he called c-AIC. Hurvich and Tsai (1989) further studied this small-sample (second order) bias adjustment which led to a criterion that is called AIC$_c$,

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

where the penalty term is multiplied by the correction factor $n/(n–K–1)$. This can be rewritten 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 (also see Sugiura 1978). AIC$_c$ merely has an additional bias correction term. If n is large with respect to K, then the second order correction is negligible and AIC should perform well. Generally, we advocate the use of AIC$_c$ when the ratio n/K is small (say < 40). In reaching a decision about the use of AIC vs. AIC$_c$, one must use the value of K for the highest dimensioned model in the set of candidates. If the ratio n/K is sufficiently large, then AIC and AIC$_c$ are similar and will tend to select the same model. One should use either AIC or AIC$_c$ consistently in a given analysis; rather than mixing the two criteria. **Unless the sample size is large with respect to the number of estimated parameters, use of AIC$_c$ is recommended.**

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." Our focus here is on a lack of independence in the data leading to overdispersion or “extra-binomial variation." Eberhardt (1978) provides a clear review of these issues in the biological sciences. For example, Canada geese (*Branta* spp.) frequently mate for life and the pair behaves almost as an individual, rather than as two independent “trials." The young of some species continue to live with the parents for a period of time, which can also cause a lack of independence of individual responses. Further reasons for overdispersion in biological systems include species whose members exist in schools or flocks. Members of such populations can be expected to have positive correlations among individuals within the group; such dependence causes overdispersion. A different type of overdispersion stems from parameter heterogeneity; that is individuals having unique parameters rather than the same parameter (such as survival probability) applying to all individuals.

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.

Given $\hat{c}$, empirical estimates of sampling variances ($var_e(\hat{\theta}_i)$) and covariances ($cov_e(\hat{\theta}_i, \hat{\theta}_j)$) can be computed by multiplying the estimates of the theoretical (model-based) variances and covariances by $\hat{c}$ (a technique that has long been used, see e.g., Finney 1971). These empirical measures of variation (i.e., $\hat{c} \cdot \hat{v}ar_e(\hat{\theta}_i)$) must be treated as having the degrees of freedom used to compute $\hat{c}$ for purposes of setting confidence limits (or testing hypotheses). Generally, quasi-likelihood adjustments (i.e., use of $\hat{c} > 1$) are made only if some reasonable lack of fit has been found (for example if the observed significance level $P \leq 0.15$ or 0.25) and the degrees of freedom $\geq$ 10, as rough guidelines.

Patterns in the goodness-of-fit statistics (Pearson χ^2 or G-statistics) might be an indication of structural problems with the model. Of course, the biology of the organism in question should provide clues as to the existence of overdispersion; one should not rely only on statistical considerations in this matter.

Principles of quasi-likelihood suggest simple modifications to AIC and AIC$_c$; we denote these modifications as (Lebreton et al. 1992),

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

and

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

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

Of course, when no overdispersion exists, $c = 1$, the formulae for QAIC and QAIC$_c$ reduce to AIC and AIC$_c$, respectively.

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."

This comment was quite insightful; of course, we might expect this from R. A. Fisher! Akaike was perhaps kind to consider AIC an extension of classical ML theory; he might just as well have said that classical likelihood theory was a special application of the more general information theory. In fact, Kullback believed in the importance of information theory as a unifying principle in statistics.

Interpreting Differences Among AIC Values

Akaike's Information Criterion (AIC) and other information theoretic methods can be used to rank the candidate models from best to worst. Often data do not support only one model as clearly best for data analysis. Instead, suppose three models are essentially tied for best, while another, larger, set of models is clearly not appropriate (either under- or over-fit). Such virtual “ties" for the best approximating model must be carefully considered and admitted. Poskitt and Tremayne (1987) discuss a “portfolio of models" that deserve final consideration. Chatfield (1995b) notes that there may be more than one model that is to be regarded as “useful." The inability to ferret out a single best model is not a defect of AIC or any other selection criterion, rather, it is an indication that the data are simply inadequate to reach such a strong inference. That is, the data are ambivalent concerning some effect or parameterization or structure.

It is perfectly reasonable that several models would serve nearly equally well in approximating a set of data. Inference must admit that there are sometimes competing models and the data do not support selecting only one. Using the Principle of Parsimony, if several models fit the data equally well, the one with the fewest parameters might be preferred; however, some consideration should be given to the other (few) competing models that are essentially tied as the best approximating model. Here the science of the matter should be fully considered. The issue of competing models is especially relevant in including model selection uncertainty into estimators of precision and model averaging.

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.*

Model Selection Uncertainty

One must keep in mind that there is often considerable uncertainty in the selection of a particular model as the “best" approximating model. The observed data are conceptualized as random variables; their values would be different if another, independent set were available. It is this “sampling variability" that results in uncertain statistical inference from the particular data set being analyzed. While we would like to make inferences that would be robust to other (hypothetical) data sets, our ability to do so is still quite limited, even with procedures such as AIC, with its cross validation properties, and with independent and identically distributed sample data. Various computer intensive, resampling methods will further improve our assessment of the uncertainty of our inferences, but it remains important to understand that proper model selection is accompanied by a substantial amount of uncertainty. Quantification of many of these issues is beyond the scope of the material here (see Burnham and Anderson 1998 for advanced methods).

AIC When Different Data Sets are to be Compared

Models can only be compared using AIC when they have been fitted to exactly the same set of data (this applies also to likelihood ratio tests). For example, if nonlinear regression model A is fitted to a data set with n = 140 observations, one cannot validly compare it with Model B when 7 outliers have been deleted, leaving only n = 133. Furthermore, AIC cannot be used to compare models where the data are ungrouped in one case (Model U) and grouped (e.g., grouped into histograms classes) in another (Model G).

Summary

The Principle of Parsimony provides a conceptual guide to model selection, while expected K-L information provides an objective criterion, based on a deep theoretical justification. AIC, 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 use 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. (1983c). On minimum information prior distributions. *Annals of the Institute of Statistical Mathematics* **35A**, 139-149.

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. (199_). General strategies for the collection and analysis of ringing data. *Bird Study* __, __-__.

Anderson, D. R., and Burnham, K. P. (199_). Understanding information criteria for selection among capture-recapture or ring recovery models. *Bird Study* __, __-__.

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**, __-__.

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.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* **52**, 345-370.

Broda, E. (1983). *Ludwig Boltzmann: man, physicist, philosopher.* (translated with L. Gay). Ox Bow Press, Woodbridge, Connecticut, USA.

Burnham, K. P., and Anderson, D. R. (1998). *Model selection and inference: a practical information theoretic approach*. Springer-Verlag, New York.

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.

Cox, D. R., and Snell, E. J. (1989). *Analysis of binary data*. 2nd Ed., Chapman and Hall, New York, NY.

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.

Eberhardt, L. L. (1978). Appraising variability in population studies. *Journal of Wildlife Management* **42,** 207-238.

Finney, D. J. (1971). *Probit analysis*. 3rd. ed. Cambridge University Press, London.

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

Fisher, R. A. (1936). Uncertain inference. *Proceedings of the American Academy of Arts and Sciences* **71,** 245-58.

Guiasu, S. (1977). *Information theory with applications*. McGraw-Hill, New York, NY.

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

Kapur, J. N., and Kesavan, H. K. (1992). *Entropy optimization principles with applications*. Academic Press, London.

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.

Nester, M. (1996). An applied statistician's creed. *Applied Statistics* **45**, 401-410.

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.

Poskitt, D. S., and Tremayne A. R. (1987). Determining a portfolio of line time series models. *Biometrika* **74**, 125-137.

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).

Tong, H. (1994). Akaike's approach can yield consistent order determination. Pages 93-103 *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.

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

Wehrl, A. (1978). General properties of entropy. *Reviews of Modern Physics* **50**, 221-260.