SOME BACKGROUND ON WHY PEOPLE IN THE EMPIRICAL SCIENCES MAY WANT TO BETTER UNDERSTAND THE *INFORMATION-THEORETIC* METHODS

Since the publication of the first edition of our book in late 1998, Ken Burnham and I have given a number of 1-, or 2-, or 4-day shortcoursesor workshops on the "Information-Theoretic" approaches to professional people with wide variety of science interests. Often, people are undecided as to whether they should take one of these shortcourses and how they might expect to benefit if they did. The purpose here is to introduce the subject in a non-technical way. This is written to allow people to gain some level of understanding as to what these approaches do and why they might be generally useful. This document is short and very incomplete, but meant as a rough, general overview of the more technical material to be covered in the short course.

The name stems from the fact that much of the deep theory underlying these approaches is based on "Kullback-Leibler information" (Kullback and Leibler 1951). This quantity is a part of *information theory* and *coding theory,* and has some interesting historical connections to Boltzmann's stunning scientific discoveries in the late 1800s concerning entropy and the Second Law of Thermodynamics (i.e., K-L information = – entropy).

K-L information (written as the function $I(f, g_i)$) is the information (I) lost when a model (g_i) is used to approximate full reality (f). Clearly, if one has R hypotheses (H_i), each represented by a model (g_i), then we would like to find the model that loses the least information about full reality. K-L information can also be viewed as a *distance,* in which case we can ask "Which hypothesis/model is *closest* to truth?" This is the so-called *model selection* problem.

In the context of interest here (model selection for data analysis and inference), it is impossible to compute K-L information directly, however, in 1973 Hirotugu Akaike derived a simple, asymptotically unbiased estimator of K-L information. This estimator has come to be known as Akaike's Information Criterion, AIC. One computes AIC for each of the R hypotheses/models and the one with the smallest value is estimated to be the best. This is a simple and very compelling concept ("select the model that is closest to truth").

The Importance of the Science and Good Data

First, the science of the issue should be firmly in place; these are not methods that will salvage poor work. Akaike believed his primary contribution was to focus people's attention on thinking, hypothesizing and modeling; if these issues are not done well, the total effort will likely be severely compromised. The objective must be clearly defined, hypotheses must be worthy of study (many seem trivial or ambiguous), and the whole effort well grounded in the science. Mathematical models must be developed to portray the science hypotheses. Good data (X) must be available, following an appropriate sampling or experimental design. [These methods are very useful for exploratory studies, but I will not make this point here.]

Multiple Working Hypotheses

Chamberlin's (1890) paper on *multiple working hypotheses* is a very effective strategy for advancing knowledge in the empirical sciences. Here, one postulates several hypotheses about a system or process of interest (call these hypotheses H_i for $i = 1, 2, ..., R$). Over 100 years ago, he did not understand *how* these hypotheses might be rigorously assessed – he only put forward the *philosophy* that multiple hypotheses should be entertained. The carefull definition of the *a priori* set of hypotheses is critical to the entire approach. A recent example of science hypotheses might be of interest at this point.

Clark and McLachlan (2003) consider the issue of forest biodiversity and define two hypotheses, each having ". . . different implications for the number and kinds of species that can coexist and the potential loss of biodiversity in the absence of speciation."

- H_1 involves stabilizing mechanisms, which include tradeoffs between species in terms of their capacities to disperse to sites where competition is weak, to exploit abundant resources effectively, and to compete for scarce resources.
- H_2 emphasizes equalizing mechanisms, because competitive exclusion of similar species is slow. Lack of ecologically relevant differences means that abundances experience random "neutral drift", with slow extinction.

The authors acknowledge that the relative importance of these 2 hypotheses is unknown, because the assumptions and predictions involve broad temporal and spatial scales. They use paleodata on more than 200 generations of 7 tree species at 8 sites in southern Ontario. [I will not go into the models they used to represent these hypotheses or the simulation approach they chose as these issues take us too far from the points of interest here. However, I might mention that I would have preferred some more hypotheses, in addition to the two they defined.]

Statistical Tests Side-Track Science for Much of the 20th Century

Chamberlin would have been disappointed by the many methods developed in the 20th century to test *null hypotheses*. That is **not** what he envisioned; the null is so often trivial or obviously false on *a priori* grounds (e.g., the population correlation coefficient between variables X and Y is 0 or the experimental addition of aluminum to an aquarium containing small fish has no effect on growth or survival). Over the past century, literally hundreds of statistical "tests" have been developed to test null hypotheses. Chamberlin envisioned ways to evaluate or contrast the support for the multiple science hypotheses; instead, tests of a usually sterile null became the unfortunate focus. These tests are not wrong in any sense, but the results from such tests seem of relatively little value (i.e., "We reject the null that was stupid in the first place, $P < 0.05$"). There are a host of problems with this traditional approach, some of these are outlined in Anderson et. al. (2000) and there are extensive websites on this old issue. In fact, controversy over null hypothesis testing among statisticians had started by the late 1930s.

A General Example

Here I generalize the example of tree diversity in southern Ontario. Consider the general issue where 3 carefully formulated science hypotheses have been proposed (H_1 H_2, and H_3): hypothesis H_1 might be the dominate thinking in the field, hypothesis H_2 might be a shift in thinking but not yet having much support, while hypothesis H_3 might be quite different and supported by only a single group. Each of these science hypotheses have been modeled (statistical expertise might be required for this important step); thus 3 models reflect the 3 hypotheses – call these models g_1, g_2, and g_3. So, we assume that hypothesis H_1 and model g_1 are essentially interchangeable; they both try to mean the same thing ($H_1 \Leftrightarrow g_1$, $H_2 \Leftrightarrow g_2$, $H_3 \Leftrightarrow g_3$).

Now, based on the data set *X* and a rigorous analysis, using either least squares or maximum likelihood methods, we can ask questions such as:

- Which science hypothesis has the most empirical support? This is the same as "which is the best model?"
- What is a proper ranking of the 3 science hypotheses? This is the same as a ranking of the models from best to worst, based on the empirical data.
- Are the first 2 hypotheses nearly tied? or are all 3 hypotheses nearly tied in terms of empirical support? Is one hypothesis far better supported by the data than the others?
- Are any of the 3 hypotheses untenable relative to the others, based on the empirical data?
- If prediction is the goal of hypothesizing and modeling, then should predictions be based on only the model estimated to be best? or should prediction be based on all the models (perhaps weighted in some way)?

These are the types of questions provided by the careful use of the information-theoretic approaches. They are not intended as cookbook recipes; instead they encourage a science focus as *a priori* hypotheses and modeling are so important. Let us consider a tutorial example and introduce (without definition) 2 quantities that are computed using information-theoretic methods (you learn this material in the first 2-3 hours of the short course). First, Δ_i is useful in ranking the hypotheses/models, whereby the best model always has $\Delta_i \equiv 0$. The other models have $\Delta_i > 0$; the larger the Δ_i value, the less plausible the model. Second, the w_i are the probability that hypothesis/model *i* is, in fact, the best model of the 3 under consideration. The Δ_i and w_i are based on the data and simple AIC values and both can be easily computed. Let the results be (for example):

i	Hypothesis	Model	Δ_i	w_i
1	H_1	g_1	3.1	0.08
2	H_2	g_2	0.0	0.92
3	H_3	g_3	16.9	0.00

Several inferences can be made here. First, the best hypothesis/model is 2, based on the data available. The full ranking (best to worst) is H_2, H_1, and H_3. Knowing some theory about the Δ_i allows one to conclude that there is very little support for H_3 (models whose Δ values are >10 have essentially no empirical support). We can ask what are the odds of hypothesis/model 3 actually being the best model; this is about 4,675 to 1 in this example. Clearly, hypothesis/model 3 lacks empirical support. How strong is the evidence that hypothesis/model 1 is better than 2? [Note, we are asking for a measure of evidence, rather than the arbitrary "significantly better" or "not significantly better."] Here, theory tells us that $\Delta_1 = 3.1$ makes this hypothesis a reasonable competitor. An "evidence ratio" can be computed; here it is 11.5. This evidence ratio (about 11) is like you and a friend buying lottery tickets: you buy 1 ticket and she buys 11 tickets. Clearly, *given that one of you wins*, the odds of her winning exceed yours; however, there is still a decent change that you will win (about 1/11). She should not want to bet her used car that she will win, when the chance of you winning is just below 10% (too risky for her). However, she might bet the whole farm if the evidence is 4,675 to 1 (as above), because her risk is very slight. So, several forms of quantitative evidence are provided by using the information-theoretic approach.

For each model, one must have either the residual sum of squares (least square analyses) or the value of the maximized log-likelihood (maximum likelihood analyses), the sample size, and the number of parameters. For these quantities (printed by essentially all computer routines for standard statistical analyses) one can easily compute the values needed for the information-theoretic approaches. Many of the examples in the book by Ken Burnham and I were done by hand with only a calculator. A spread sheet would be handy for problems with many hypotheses/models.

So, to summarize so far, the information-theoretic approaches allow a quantitative strength of evidence to be computed and interpreted with respect to the science hypotheses of interest. This is a useful set of tools, whether one has only 2 hypotheses/models or 32 or more. Chamberlin's concept was that the set of multiple working hypotheses would evolve over time. As more experiments or sampling studies were conducted and the data analyzed, some hypotheses would be dropped (e.g., Δ_i>10 or 12) but others would be added and stand ready for evaluation with new data.

Multimodel Inference – Model Averaging

There are a raft of theoretical advantages in basing formal statistical inference on more than one model. Nearly all of the statistical literature and thinking has been grounded on the idea than you somehow get a model; inference is then based on this single model. Rather than sticking one's neck out and basing inference (e.g., predictions) on the model estimated to be best; it is often better to make formal inference based on all the models. This is done quite simply by a weighted average, where the weights are the model probabilities (w_i). Consider predicting the value of a response variable $\hat{Y}$ (Y might be fish biomass) from each of the 3 hypothesis/models in the example above. The model averaged estimate of fish biomass is just $\hat{Y} = \sum w_i \hat{Y}_i$, where i = 1, 2, 3 and indexes the 3 models. In this example, the contribution of hypothesis/model 3 is essentially zero, with hypothesis/model 2 carrying 92% of the weight.

Multimodel Inference – Relative Importance of Predictor Variables

Often one has *r* predictor variables and would like a ranking of their relative importance. This issue is complicated by the fact that measurements of the variables might in feet or inches, or meters, or acres or square kilometers, or pounds or kilograms, etc. Another complication is that X_i and X_j might be highly correlated, often giving a false sense of importance to one variable just because it happens to be correlated with another. Simple information-theoretic methods allow a ranking of the relative importance of predictor variables and this is often of interest, particularly in exploratory studies.

Multimodel Inference – Honest Estimates of Precision

Many commonly used methods provide an estimate of the precision of some estimated parameter (often the sampling variance or standard error, coefficient of variation, or a confidence interval). These are computed *assuming* a model; that is, they are conditional on a model. In the real world, we do some analysis to select a model (e.g., stepwise regression – however this is a surprisingly poor approach, although very commonly used) and have only an *estimate* as to which model should be used. There is uncertainty in the data-based selection of a model, but this is not reflected in the usual estimates of standard error (Breimann called this a quiet scandal). Information-theoretic approaches allow estimates of precision to include a variance component for "model selection uncertainty." This is trivial to compute and understand (with some background given during the short course).

Model averaging, relative importance of variables, and incorporating model selection uncertainty into estimates of precision are forms of multimodel inference; the central theme of these short courses.

Theory and Application

The theory underlying the information-theoretic methods is very deep (see Chapter 7 of our second edition or the short paper by Kullback and Leibler), but the *application* of the general approach is simple. The main formulae are summarized on a single laminated reference sheet that is distributed at short courses. In addition, the approach applies to a very broad class of science problems and model types. A reading of Chapter 8 of the second edition of Burnham and Anderson (2002) might be helpful to those wanting still more insights into this class of methods.

In contrast, a biologist trying to test a statistical null hypothesis is faced with an array of parametric and nonparametric tests; this may be some type of randomization procedure that is quite computer intensive (several of these tests do not even give the estimated effect size as part of the computer output). Tests may be considered 1- or 2-tailed and computed test statistics can be asymptotically distributed as χ^2, *z, t, F*, etc., etc. Little generality exists, forcing the scientist or manager to be familiar with dozens of different test procedures and many of these cannot be computed by hand.

More on Null Hypothesis Testing

The most fundamental problem with null hypothesis testing is that nearly all nulls are uninteresting or trivial (many are actually stupid); thus, the results do not allow increased understanding or provide new insights. An example is taken from a recent issue of *Ecology* where the authors define the null hypothesis that species diversity (H') is constant over time. To me, this null seems absurd; however, the authors propose a computer intensive test procedure. Finally, they illustrate the approach with data on 8 species of dinosaurs and their species diversity over *geologic* time! I must ask how science is advanced by rejecting this null hypothesis (i.e., as biologists, how can we consider this null to be at all plausible?). Tests of null hypotheses are usually fairly uninformative, but they have been nearly mandatory for many journals and may sometimes give the false impression of "good" science.

Bayesian Methods

Methods based on Bayes' theorem are on the increase in recent years and these have merit for biologists working closely with a good PhD-level statistician with a background in such methods. While I am not a Bayesian, I support the general approach (particularly with "non-informative" priors). However, I find that the technical level is beyond nearly all subject matter scientists, the computational issues are still daunting, and they have not reached any agreement on the model selection problem. [Bayesians also have a low regard for null hypothesis testing.] Still, I have a positive attitude toward Bayesian methods; however, my shortcourses focus on the information-theoretic approaches.

Go to *http://aicanderson*1*.home.comcast.net* for more technical information. Several related reprints can be found at *www.cnr.colostate.edu/~anderson.sel_reprints.html*

Literature Citations

Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64, 912-923.

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

Burnham, K. P., and D. R. Anderson. 2002. *Model selection and multimodel inference: a practical information-theoretic approach*. 2nd Ed. Springer-Verlag, New York, NY. ISBN 0-387-95364-7.

Chamberlin, T. C. 1890. The method of multiple working hypotheses. Science 15, 93-98.

Clark, J. S., and J. S. McLachlan. 2003. Stability of forest biodiversity. Science 423, 635-638.

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

David R. Anderson
November 1, 2003
www.cnr.colostate.edu/~anderson