Behav Ecol Sociobiol (2011) 65:23–35
DOI 10.1007/s00265-010-1029-6

REVIEW

AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons

Kenneth P. Burnham · David R. Anderson · Kathryn P. Huyvaert

Received: 15 May 2009 / Revised: 17 July 2010 / Accepted: 28 July 2010 / Published online: 18 August 2010

Abstract We briefly outline the information-theoretic (I-T) approaches to valid inference including a review of some simple methods for making formal inference from all the hypotheses in the model set (multimodel inference). The I-T approaches can replace the usual *t* tests and ANOVA tables that are so inferentially limited, but still commonly used. The I-T methods are easy to compute and understand and provide formal measures of the strength of evidence for both the null and alternative hypotheses, given the data. We give an example to highlight the importance of deriving alternative hypotheses and representing these as probability models. Fifteen technical issues are addressed to clarify various points that have appeared incorrectly in the recent literature. We offer several remarks regarding the future of empirical science and data analysis under an I-T framework.

Keywords AIC · Evidence · Kullback–Leibler information · Model averaging · Model likelihoods · Model probabilities · Model selection · Multimodel inference

Communicated by L. Garamszegi

This contribution is part of the Special Issue "Model selection, multimodel inference and information-theoretic approaches in behavioral ecology" (see Garamszegi 2010).

K. P. Burnham (✉) · D. R. Anderson
Colorado Cooperative Fish and Wildlife Research Unit,
Colorado State University,
Fort Collins, CO 80523, USA
e-mail: kenb@lamar.colostate.edu

K. P. Huyvaert
Department of Fish, Wildlife, and Conservation Biology,
Colorado State University,
Fort Collins, CO 80523, USA

Introduction

The broad theoretical concepts of information and entropy provide the basis for a new paradigm for empirical science. Good science is strategic and an excellent strategy begins with Chamberlin's (1890) "multiple working hypotheses." This principle encourages hard thinking to identify the alternative science hypotheses: H_1, H_2, ..., H_R. The careful identification of this *a priori* set of hypotheses is very important and is at the center of the science issue (see Elliott and Brook 2007). In the past, it has been standard practice to define a single alternative hypothesis and a null hypothesis but this procedure can be improved upon. We suggest that investigators make a major effort to think hard about the science question and then define several plausible alternative hypotheses—inferences are conditional on this set of alternatives. For example, Hall (2004) provides a nice set of alternatives for explanations of avian duetting; the next step would be to gather appropriate data and evaluate the strength of evidence for the hypothesized models.

Chamberlin said little about how one might evaluate the relative worth of these alternatives other than wanting a "measure of probability on one side or the other" (Chamberlin 1890, p. 758). Perhaps he would have been content if there existed a simple way to *rank* the alternatives. New methods, based on Kullback–Leibler (K-L) information, provide a formal relative *strength of evidence* for each of the alternative hypotheses. Obtaining quantitative measures of the strength of evidence for each hypothesis (H_i) represents the fundamental methodological issue in empirical science and, appropriately, can be considered an important advance for the life sciences, including behavioral ecology.

Quantitative measures of the strength of evidence are central to empirical science; however, one is hard pressed to

find this word or the concept in books on applied or theoretical statistics. The concept of strength of evidence seems almost "new" in the life and social sciences. Traditional methods have focused on "testing" null hypotheses based on test statistics and their associated *P* values. From the *P* value comes an arbitrary judgment concerning "statistical significance" and dichotomous ruling about the rejection of, or failure to reject, the null hypothesis. For several reasons, *P* values do not constitute a basis for formal evidence (see Royall 1997). The new I-T methods are not a test in any sense; rather they represent a very different methodology for empirical science. The I-T methods provide a formal, fundamentally sound, approach of developing an *a priori* set of hypotheses and then a quantification of the data-based evidence for, and ranking of, each hypothesis. This is followed by interpretation of the results in the face of model selection uncertainty and this is one aspect of multimodel inference.

The twentieth century brought quantification and mathematical models into the process of science. The quantification of the set of alternative hypotheses provides many important advantages. Thus, it is necessary to derive a mathematical model (g) for each of the R hypotheses: g_1, g_2, ..., g_R. Ideally, there is a one-to-one mapping of the R hypotheses with their models. This allows one to treat each hypothesis and its model as synonymous. The quantification also brings in a large measure of rigor in the process of science.

A general theory of "information" developed rapidly during the 1940s. The concept of "information" was quantified and this provided a series of enormous breakthroughs affecting modern society (see Hobson and Cheng 1973, Guiasu 1977, Soofi 1994, Jessop 1995, and Cover and Thomas 2006 for background). Kullback and Leibler (1951) worked in the war effort to develop ways to break codes and their results are most relevant to this review. At the same time, Shannon (1948) was developing a mathematical theory of communication. Today society is still benefiting from this broad theory (e.g., cell phone and GPS technology). Information theory was linked to statistical theory in the early 1970s by Hirotugu Akaike in a series of groundbreaking papers (e.g., Akaike 1973, 1974, 1977, 1981a, b, 1983a, b).

Developments from information and statistical theory allow a quantification of the strength of evidence for each of the alternative hypotheses in the set. These results have lead to several powerful new approaches to empirical science. Our objectives in this paper are to briefly review the fundamentals of the "information-theoretic" or "I-T" approaches and to show how these easily extend to making formal inferences from all the models in the set (i.e., multimodel inference). This is followed by an ecological example and clarification of 15 technical issues that have arisen in the literature. We then offer some contrasts between the traditional null hypothesis testing approach and the array of I-T approaches. We conclude with some predictions about the future of empirical science under an I-T framework.

This paper is meant to be an overview of the central aspects of the I-T approaches. However, we do not provide much of the underlying theory or discuss many of the finer points as these issues are given in detail by Burnham and Anderson (2002) and explained in the brief tutorial book by Anderson (2008). We must assume that the reader has some knowledge of and, perhaps, experience with the I-T approaches and a reasonable background in statistical concepts (e.g., least squares "regression," expected values, measures of precision, goodness-of-fit assessment). Without this background, not all points can be easily understood; this is not meant to be a first introduction to this subject (such as Anderson 2008 is), but is intended to provide a brief overview and references for additional study (see papers by Garamszegi 2010, and Symonds and Moussalli 2010, for introductory material).

Fundamentals

The new class of approaches is called "information-theoretic" because it is based on K-L information (Burnham and Anderson 1992, 2001, 2002, 2004). We give an overview of the main issues in the material to follow. K-L information as applied here represents the *information lost* when model g_i is used to approximate full reality (f). Another view of this is the *distance* between model g_i and full reality. In either case, it seems compelling that one would want to select the model in the set of R models that minimizes K-L information loss. That is, we want the model from within the model set that loses the least information about full reality, hence, the model that is closest to full reality in the current model set (Fig. 1). Formally, K-L information can be expressed as a difference between two statistical expectations (Burnham and Anderson 2002, p. 58 and Anderson 2008, p. 52–7). The first such expectation cannot be computed or estimated but is constant across models and can be removed. The relevant term is the second expectation, $E[\log(g(x|\theta))]$, where E is the expectation operator, log is the natural logarithm, x represents the response variable to be predicted by the model (x represents hypothetical data), and θ represents a vector of unknown parameters. This second term also cannot be computed or estimated.

However, in a famous paper published in 1973, Akaike found that if a second expectation was taken over an estimated θ then that quantity could be estimated and this result provided the link between K-L information and the

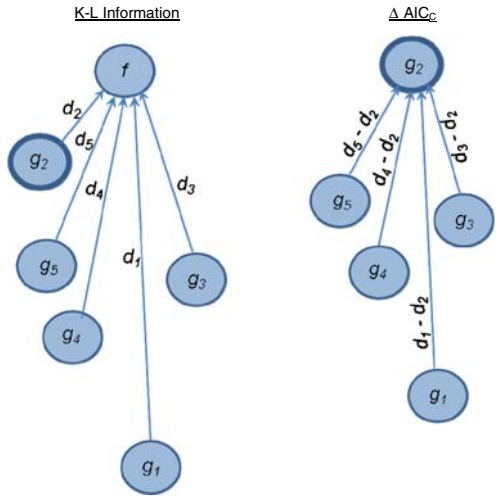

Fig. 1 Kullback–Leibler information is shown (at left) as the distances (d_i) between full reality (f) and the various models (g_i). The Δ values (right) provide the estimated distance of the various models to the best model (in this case, model g_2). These values are on the scale of *information* irrespective of the scale of measurement or type of data. The Δ values are simple to compute, allow a quick ranking of the models, and are the key to multimodel inference

maximized log-likelihood (a fundamental quantity in mathematical statistics, written as $\log(\mathcal{L})$). The concept of likelihood is fundamental to statistical theory (see Edwards, 1992, Azzalini 1996, Severini 2000, and Pawitan 2001). Akaike's key finding focused on the double expectation,

$$\mathrm{E}\left\{\mathrm{E}\left[\log\left(\mathrm{g}\left(x\middle|\widehat{\theta}(y)\right)\right)\right]\right\}$$

where y represents data and $\widehat{\theta}\left(=\widehat{\theta}(\mathrm{y})\right)$ is the vector of parameter estimates based on these data. Akaike found that for large sample sizes (n) this double expectation can be estimated very simply as $\log(\mathcal{L}) - K$, where K is a correction for asymptotic bias and is merely the total number of estimable parameters in the model. That is,

$$\mathrm{E}\left\{\mathrm{E}\left[\log\left(\mathrm{g}\left(x|\widehat{\theta}(y)\right)\right)\right]\right\} = \log(\mathcal{L}) - K.$$

Akaike multiplied both terms by −2 to get his $\mathrm{AIC} = -2\log(\mathcal{L}) + 2K$. The term $-2\log(\mathcal{L})$ is well known among statisticians as the "deviance;" Akaike no doubt thought of AIC as simply "deviance plus $2K$" and many software packages provide the deviance as part of the standard output.

Operationally, one computes AIC for each of the R models and selects the model with the smallest AIC value as "best." Such a model is "best" in the sense of minimizing K-L information loss. Full details of the derivation from K-L information to AIC are given in Burnham and Anderson (2002, chapter 7), while Anderson (2008, chapter 3) provides a simplified sketch of this derivation.

A second order bias correction for AIC was derived by Sugiura (1978) and Hurvich and Tsai (1989) and is important to use in practice, particularly when sample sizes are small as often applies to behavioral studies. This criterion is denoted as AICc to make it distinct from AIC,

$$\mathrm{AICc} = \mathrm{AIC} + (2K(K+1))/(n-K-1).$$

As sample size (n) increases, AICc converges to AIC.

In the case of ordinary least squares regression or analysis of variance,

$$\log(\mathcal{L}) = -(n/2)\log(\mathrm{RSS}/n),$$

thus

$$\mathrm{AICc} = n\log(\mathrm{RSS}/n) + 2K + (2K(K+1))/(n-K-1),$$

where RSS denotes the residual sum of squares from the fitted model.

AICc implicitly has additive unknown constants that do not depend upon the fitted model. Thus, it is the AICc differences, Δ values or simply Δs, that are pivotal for ranking the models according to K-L information loss (or distance).

$$\Delta_i = \mathrm{AICc}_i - \mathrm{AICc}_{\min}, \quad \text{for } i = 1, 2, \ldots, R.$$

Here, $\mathrm{AICc}_{\min}$ denotes the minimum of the AICc values for the R models. These Δ values are on a continuous scale of *information* and are interpretable regardless of the measurement scale and whether the data are continuous or discrete or categorical. Models with Δ values above about 9–11 have relatively little support (Table 2, and Fig. 2); that is, these models lose too much information about full reality relative to some other models in the set. Going further, Δ values greater than, say, 20 have essentially no empirical support. The Δ values are the key to I-T approaches and corresponding multimodel inference.

The procedure is simple to both understand and compute. One computes AICc and Δ values for each hypothesis and selects the one with the smallest information loss or smallest distance from full reality as the best hypothesis and obtains a ranking of the rest.

The (relative) likelihood of each model i, given the data, $\mathcal{L}(\mathrm{g}_i|\mathrm{data})$ can be denoted as just ℓ_i These model like-

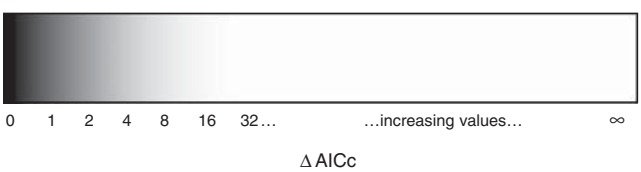

Fig. 2 Plausible hypotheses are identified by a narrow region in the continuum where Δ< perhaps four to seven (*black and dark grey*). The evidence in the light grey area is inconclusive and value judgments for hypotheses in this region are equivocal. Implausible models are shown in *white*, Δ> about 14

lihoods provide a formal strength of evidence for each of the models in the set and are easy to compute,

$$\ell_i = \mathcal{L}(g_i|\text{data}) = \exp(-(1/2)\Delta_i).$$

These quantities allow evidentiary statements such as, for example, given the data, "hypothesis H_4 is 22 times more likely than H_2." Such inferences are very useful (e.g., "Model g_3 is 1,178 times better supported by the data than model g_6").

The probability of each model g_i, given the data and the R models, is also simple to compute as a measure of strength of evidence,

$$w_i = \text{Prob}\{\text{model } g_i|\text{data}\} = \ell_i / \sum_{j=1}^{R} = \ell_j.$$

These quantities would have been the envy of Chamberlin and allow statements such as "the probability of H_4 is 0.78, while the probability of H_2 is 0.015." Clearly, the data support H_4 in this example.

Finally, one can take ratios of either the model likelihoods or model probabilities for any two models i and j to compute an "evidence ratio." In the example just above, the evidence ratio for H_4 versus H_2 is 0.78/0.015=52: the empirical support for H_4 is 52 times that of H_2. This evidence might be judged to be "strong". The quantitative evidence is represented by the model likelihoods, model probabilities, and evidence ratios; these are the science results. Then a value judgment is made as the results are interpreted and qualified. Such value judgments attempt to explain the science result. The word "significant" is to be avoided as it implies the older approaches and implies a dichotomy (reject or not) that is not appropriate.

The Δ values for any given model are linked to the evidence ratio for the best model as $\exp\{-(1/2)\Delta\}$ and a sample of these are summarized in Table 1. For example, if a model in question has a Δ value of 11, then its evidence ratio compared to the best model is approximately 245. That is, the evidence is 245 times stronger for the best model relative to the model in question. People might often judge this evidence to be very strong, other people might choose another word; however, both judgments are based on the same quantitative evidence, an evidence ratio of 245 to 1.

In summary, the evidence for each model in the set can be quantified using model likelihoods, model probabilities, and evidence ratios. Note, however, that the Δ values are central as they are on the scale of information. The fact that information can be quantified has proven to be very useful (Fig. 2).

Table 1 A summary of the strength of evidence for the best model versus model j in terms of its Δ value

Δ_j	Evidence ratio
2	2.7
4	7.4
6	20.1
8	54.6
9	90.0
10	148.4
11	244
12	403
13	665
14	1,097
15	1,808
20	22,026
50	72 billion

Multimodel inference

In many cases there is substantial model selection uncertainty; the analyst is uncertain as to which is actually the K-L "best" model. This uncertainty is quantified by the model probabilities (e.g., the best model has only probability 0.47). Often, a particular model is *estimated* to be the best of those in the model set; however, there may be substantial uncertainty over this selection. In addition, there is usually information in the second, third, fourth, and other models that is not captured by the best model. Thus, basing inference only on the model *estimated* to be the K-L best represents poor practice. This thinking leads to the concept that inferences should often be based on all the models in the *a priori* set, not just the one *estimated* to be best.

The first approach is called model averaging (see Hoeting et al. 1999 for background and discussion from a Bayesian viewpoint) and this can be best understood from the viewpoint of prediction. Let $\widehat{Y}_i$ be the predicted value from the ith model, where i=1, 2, … R, given fixed values of the predictor variables. A model-averaged prediction can be computed as a weighted mean where the weights are the model probabilities,

$$\widehat{\overline{Y}} = \sum_{i=1}^{R} w_i \widehat{Y}_i.$$

Other approaches to multimodel inference include simple ways to compute measures of precision that include a variance component for model selection uncertainty,

$$\text{var}\left(\widehat{\overline{Y}}\right) = \sum_{i=1}^{R} w_i \left\{ \text{var}\left(\widehat{Y}_i|g_i\right) + \left(\widehat{Y}_i - \widehat{\overline{Y}}_i\right)^2 \right\}.$$

There are approaches to averaging model parameters; these are outlined in Burnham and Anderson (2002, pp. 150–153) and Anderson (2008, section 5.1). General methods also exist for ranking the relative importance of

Table 2 A selection of ecological factors that could be predictors of the occurrence of extra-pair paternity in birds

Factors	Link to extra-pair paternity	Example(s)
Male age (X_1)	EPY more likely in nests with younger males	Schmoll et al. 2007
	EPY more likely in nests with older males	Perreault et al. 1997
Male body size/condition (X_{2i}, X_{2ii})	EPY less likely in nests with larger males or males in better 'condition'	Yezerinac and Weatherhead 1997 Currie et al. 1999
Genetic similarity of social mates (X_3)	EPY more likely when social mates are genetically similar (female have EPCs with less similar males to avoid inbreeding)	Eimes et al. 2005; but see Schmoll et al. 2005
Food availability (X_4)	EPY more likely in nests on territories with high food abundance	Hoi-Leitner et al. 1999
Dominance status of male (X_5)	EPY more likely in nests with low-status males	Mennill et al. 2004
Female prospects for other mating opportunities (X_6)	EPY more likely as female opportunities for EPC increase	Brylawski and Whittingham 2004
Territory quality (X_7)	EPY more likely in nests on low quality territories	Rubinstein 2007

predictor variables in large regression or discriminant function analyses, and for computing confidence sets on models (details are provided in Burnham and Anderson 2002, chapter 4).

Empirical science in the twenty-first century will increasingly rely on multimodel inference. Models in the life sciences are nearly always oversimplified and it is not reasonable to make inference based on only the one model *estimated* to have been the best (i.e., "best" in the sense of K-L information). Rather, there are a host of advantages in making inference from a weighted combination of results from all the models in the set. Estimates of precision must account for model selection uncertainty or else confidence coverage will often be well below the nominal level (given a frequentist's interpretation of confidence limits). Appropriate estimates of precision that account for model selection uncertainty are easily done in the I-T framework.

After the data have been carefully analyzed, one can review the quantitative evidence and consider some qualifying value judgments to help understand and interpret the evidence. At that point new hypotheses are often formulated, based on the results from the *a priori* efforts. We encourage such further analysis of the existing data; however, the new results stem from *post hoc* analyses and the process must be admitted and the results treated with additional caution. We encourage consideration of *post hoc* alternatives and analyses but they should follow the *a priori* steps and be fully explained in publications.

Science can move at a fast pace if the set of alternatives "evolves" by: (1) discarding hypotheses judged to be implausible, (2) refining hypotheses that seemed plausible, and (3) adding new and perhaps more sophisticated alternative hypotheses by more thinking and synthesis (see e.g., Platt 1964). Of course, new data are required at each step of this evolution. Science is always "asking for more."

A hypothetical example

Here we present a hypothetical example meant to present a simplified overview of the process through hypothesis generation to data collection to analysis in an I-T framework. We draw on the extensive literature concerning extra-pair paternity (EPP) in birds to illustrate the I-T approach. Based on observations of a small number of subfamilies of birds, Lack (1968) estimated that 90% of bird species' mating system was "monogamous," which describes an exclusive social and sexual "pair bond" of one female and one male over some period of time. Since Lack's work, researchers have documented extra-pair copulations (EPCs), in a wide range of species, from bluebirds to albatrosses. The advent of molecular techniques has lead to the recognition that these EPCs can lead to extra-pair fertilizations with EPP accounting for, on average, 11% of young produced in socially monogamous birds (see Griffith et al. 2002 for a recent review).

Studies dedicated to understanding EPP in birds have examined a diversity of behavioral and ecological correlates to parentage at both inter- and intraspecific levels (Griffith et al. 2002, Westneat and Stewart 2003, and Neudorf 2004 provide overviews); these factors are not typically considered in concert as multiple competing hypotheses, and often no formal ranking or strength of evidence for the various hypotheses is presented. Here, we recast the science question about extra-pair paternity to present an example of how workers in the behavioral sciences might undertake a research program using an I-T approach.

Imagine a socially monogamous bird species where male care is essential; without the male's efforts at incubation, provisioning the young, or other caretaking, reproductive success for a breeding bout is zero. From molecular work on samples collected during a pilot field season, we also know that about 20% of young are extra-pair so that a

portion of males are not the genetic fathers of the chicks they raise each breeding bout (e.g., per season, per clutch). Which measurable ecological conditions (covariates, to be denoted "X") predict the occurrence of extra-pair young (EPY) in nests of a sample of our example species?

Ideally the data for this effort have not yet been collected and we can place considerable effort into hard thinking (this stage may take many months) to develop an ecologically meaningful *a priori* set of models—hypotheses predicting how and when EPP occurs in nests of our example species. In addition, we might consult the literature and experts in the field to broaden the catalog of alternative hypotheses (Table 2). In light of pilot data and additional careful consideration, we may omit particular explanations from the model set because they are not plausible or feasible or we may consider additive models or interactions of various predictors (see also Dochtermann and Jenkins 2010, for more on developing the *a priori* hypothesis set). In this example, we have developed a model set of 14 ecologically plausible hypotheses for the occurrence of EPY in nests of our species (Table 3). We also include the intercept-only model as a baseline for comparison. As an upper limit, *R*, the number of models, should be less than the expected sample size.

In this simplified example, we begin with the set of factors that we found to be important in the literature, our own experience, and in talking with expert colleagues (Table 2). After careful thinking (and some heated discussion among us), we omitted male age, genetic similarity, and female prospects from our list. For example, we excluded male age because a demographic study that a colleague conducted showed that most of the individuals from our example species of bird reproduce at the age of 2 years, leaving very few at other ages with which to evaluate the importance of age on the probability of EPY. Genetic similarity was excluded for this first analysis because we found during our pilot season that these birds are often secretive during copulations making it difficult to identify a candidate set of males (e.g., EPC partners for females) from whom to try to assign paternity (note that there are several excellent cases in the literature where it has been possible to assign paternity [e.g., Richardson and Burke 2001]). Similarly, we omitted female prospects because of our concerns in being certain we could accurately measure opportunities for EPCs (i.e., available males that are not the social mate). Additional information from the first season's data collection and modeling effort or from ancillary studies may prompt us to reconsider these variables again later. Thus, the set of science hypotheses may "evolve" and analyses of the new evolved model set can be performed using newly collected data.

We included these four factors of interest in linear models where the covariate is related to the logit of the response variable, parentage of the youngster. We then considered additive models of pairs of predictor variables. For example, after more thought, territory quality and food availability seemed similarly important so both of those covariates were related to the response variable in an additive generalized linear model. In considering data from our pilot study, we found that male status and territory quality were very tightly correlated, perhaps because the highest ranking males could secure the best quality territories. We omitted additive models including both of these variables together in the same model because we think they are redundant (see e.g., Freckleton 2010) Lastly, we considered interactions between pairs of the variables and added to our model set interactive models of body size with male dominance status, body size with territory quality, and food availability with territory quality (Table 3). For example, the thinking for the model with the interaction of food availability with territory quality was that the effect food availability has on the probability of having an EPY in the nest differs for nests from different territory qualities. While complex, we may want to examine a three-way interaction in subsequent years if, for example, food is very abundant in the first year and not others, perhaps because of unusually high levels of precipitation in the first year.

With the model set in hand, we recommend further consulting with knowledgeable colleagues, the literature, and a statistician before collecting data to clarify issues of

Table 3 A model set for the example examining ecological factors and extra-pair paternity in a hypothetical bird species

Model description	Model notation
Male body size ('body')	$\beta_0+\beta_{2i}X_{2i}$
Food availability ('food')	$\beta_0+\beta_4X_4$
Male dominance ('status')	$\beta_0+\beta_5X_5$
Territory quality ('territory')	$\beta_0+\beta_7X_7$
Body+food	$\beta_0+\beta_{2i}X_{2i}+\beta_4X_4$
Body+status	$\beta_0+\beta_{2i}X_{2i}+\beta_5X_5$
Body+territory	$\beta_0+\beta_{2i}X_{2i}+\beta_7X_7$
Food+status	$\beta_0+\beta_4X_4+\beta_5X_5$
Food+territory	$\beta_0+\beta_4X_4+\beta_7X_7$
Body+food+status	$\beta_0+\beta_{2i}X_{2i}+\beta_4X_4+\beta_5X_5$
Body+food+territory	$\beta_0+\beta_{2i}X_{2i}+\beta_4X_4+\beta_7X_7$
Body×status	$\beta_0+\beta_{2i}X_{2i}+\beta_5X_5+\beta_{2i,5}(X_{2i}{*}X_5)$
Body×territory	$\beta_0+\beta_{2i}X_{2i}+\beta_7X_7+\beta_{2i,7}(X_{2i}{*}X_7)$
Food×territory	$\beta_0+\beta_4X_4+\beta_7X_7+\beta_{4,7}(X_4{*}X_7)$
Intercept only	β_0

The models link the hypothesized predictor variables from Table 2 with the probability of having an extra-pair young in the nest. Each model is a representation of the biological hypothesis of interest. Models are of the form, logit [Prob(EPY)]=$\beta_0+\beta_zX_z$, but the notation is truncated here to list the intercept and response variables and their relationship to each other, if applicable

study and sampling design so that the data collected are those most appropriate for addressing the hypotheses of interest. This is particularly important given the methodological and logistical challenges that studying birds in the wild presents (e.g., secretiveness, mobility, difficulty in capture, environmental variability, etc.). After this hard thinking and careful planning, relevant data to collect for this example study would include measures of food availability such as numbers of insects trapped per territory, measures of territory quality, and behavioral observations to establish male dominance status. All adults and young in the study should be measured (for condition or body size metrics) and small samples of blood obtained for DNA-based molecular determinations of parentage and storage for subsequent additional analyses. Importantly, measures for all of the variables of interest are needed for each case that will be considered in the analysis data set (i.e., the same data will be included for each model).

In terms of analysis in this example, our predictor variables are our metrics of territory quality, food availability, male dominance status, and body size. The response variable is proportion of the brood that is extra-pair. For simplicity, assume our example species lays clutches of only one egg, so the response, Y, will be binomial, $Y=1$ for an EPY in the nest and $Y=0$ for a within-pair young in the nest. The parameter of interest is the expected value of Y, which is equivalent to the probability $Y=1$.

A good analysis strategy will be to use generalized linear models (McCullagh and Nelder 1989) with a logit link (i.e., logistic regression) which is straightforward to implement in many software packages (e.g., SAS, Statistica, R) and is a standard, commonly used data analysis method. Other model forms (e.g., quadratic or other nonlinear shapes) and link functions (e.g., the complementary log–log link) might also apply and these approaches can be applied when the brood size is larger than 1 and thus the response variable is the proportion of the brood that was extra-pair.

If values of AICc are not supplied by the software, then it is a simple matter to calculate the AICcs and Δs from the maximized log-likelihood (or "deviance") as outlined above. Model ranks, model probabilities, and evidence ratios can then be calculated. The focus should be on the alternative science hypotheses and models that carefully reflect these. Given hard thinking and relevant data, the computations are quite easy (once the model fit has been accomplished.)

Technical issues

In general, we believe that the application of these new approaches in the life sciences has gone fairly well in a relatively short time period. However, we also note several recent methodological papers published that contain some misinformation. In this section, we will comment on a number of technical issues where emphasis and clarification might be helpful. The points below are not in any particular order.

1. *Hard thinking*. Steidl (2007) notes that the I-T approaches encourage, if not require, a person to think hard about alternatives. In the past it has been all too easy for a person to start with an interesting research hypothesis and then produce a competing, but usually trivial, null hypothesis. Then, it is only the null that is the subject of the test. Thus, often the uninteresting null is "rejected" and support is thrown to the original hypothesis, but only by default. The original science hypothesis is never tested. However, if there was little or no *a priori* belief in the null, what has been learned by its rejection?
2. *Stepwise AIC*. Stepwise regression is a very poor procedure, although well known and often taught and used (Whittingham et al. 2006). The technical reasons for its poor performance are many, but include the "multiple testing problem." The analyst does not even know the second best model when using one of the test-based subset regression methods (Mundry and Numm 2009). There are no model likelihoods or model probabilities under this approach. One cannot model average predictions or model parameters using this traditional method. Finally, estimates of precision cannot include model selection uncertainty; thus confidence intervals will be too narrow and coverage will not be at the nominal level.

   Some computer software now implement a "stepwise AIC" procedure that tries to avoid some of the worst features of the traditional stepwise testing procedure. We cannot recommend this approach as there is no theory underlying the approach and its properties are unknown. Finally, stepwise AIC bypasses the hard thinking step that is so important in empirical science. Instead, it is usually a strategy of pretending to run all possible models while, in fact, only a relatively few models are actually evaluated.
3. $\Delta>2$ *Rule*. Some of the early literature suggested that models were poor (relative to the best model), and might be dismissed if they had $\Delta>2$. This arbitrary cutoff rule is now known to be poor, in general. Models where Δ is in the 2–7 range have some support and should rarely be dismissed (see Fig. 2). Inference can be better based on the model likelihoods, probabilities, and evidence ratios and, in general, based on all the models in the set. From these quantitative measures one can then assign their own value judgment if they wish.

4. *True models*. A number of model selection methods, and much research on model selection methods, rests on the existence of a "true model" and that such a model is in the *a priori* set. The simplest interpretation of the "true model" is that the real data were actually *generated* by this unknown model; alternatively, it is a model that expresses full reality in all its aspects. Models are only approximations, by definition if nothing else. Surely no one would say full reality is a model! Models are like maps, they can be useful at various scales, but are never completely "true." The I-T approaches make no use of any such "true model" but, instead, rely on estimates of the distances of different models from full reality.

   Related to this issue is the Bayesian Information Criterion (BIC) which has been touted as being "consistent." Here the notion is that the criterion identifies the true model with probability 1 as sample size goes to infinity (Schwarz 1978). Of course, if the true model is not in the set under consideration, there is nothing to be consistent for. There are a host of reasons why BIC is a poor criterion (Burnham and Anderson 2004); we believe it should not be used with real data. Unfortunately, several computer software packages provide BIC in the output.
5. *Mixing analysis paradigms*. A common problem is where authors use null hypothesis testing methods and information-theoretic methods in the same analysis. This has been advocated in the literature and we strongly advise against it. Often people will rank the hypotheses and the associated models using AICc and then "test" to see if the best model is "significantly better" than the second best model. It is not clear why this might be interesting but it arises fairly often. We make two points. Firstly, if this is a question of interest, a simple evidence ratio of the two best models is far more informative and theoretically sound. Secondly, the theory is lacking for a traditional test because one has no idea of the distribution of the test statistic under the null (the null here is the second best model) because data analysis has been done to rank the top two models (using AICc). We strongly recommend using one paradigm or the other, but not mixing them in the same overall analysis.
6. *The meaning of model probabilities*. We have seen some confusing definitions of the I-T model probabilities, w_i, in recent publications. The correct interpretation is simple. Firstly, it must be clear that one of the R models is, in fact, the theoretically best model in a K-L information sense. Of course, the analyst does not know which of the models in the set is actually best, given only a single data set. We can *estimate* which model is best and the model probabilities quantify the probability of each model in the set being that best model. If, for example, g_3 is in fact the theoretically K-L best model and the Prob (g_3 |data)= 0.99 one can rest assured that g_3 really is the best model, given the data. Taking the example further, if Prob (g_3 |data)=0.43, then the analyst must realize that there is considerable uncertainty in the data-based selection of the best model. That is, if one had a replicate data set from the same system, it may well be that some other model would be *estimated* to be best. Here, there is a lot of uncertainty as to which model is actually the K-L best model. Model probabilities under an I-T framework have no connection to a supposed "true" model that is assumed to be in the model set.
7. *The meaning of a P value*. The definition of a P value might seem strained. One starts with experimental data and then computes a test statistic that has a known distribution by design (e.g., t or F or z or χ^2). A P value is then the probability that a test statistic would be as large as, or larger than, the actual computed test statistic, given the null. It is a "tail probability" and for this reason (there are others) P values are not evidence (Royall 1997). People often want to "redefine" such P values to be the probability of the null, given the data—this is seriously wrong (see e.g. Sellke et al. 2001).
8. *AIC only for two models.* We have seen a recent book that states that AICc can only be used to compare two models. This is simply incorrect. Strengths of the I-T approaches are the ability to deal effectively with complex problems and the ability to make formal inference from many models.
9. *Nested and non-nested models*. We have seen papers that claim that AICc can only be used for nested models and papers claiming it should be used only for non-nested models. Neither of these claims are correct.
10. *Why not just use the global model?* It has been argued that one should make inference from a model with all the factors thought to be important (i.e., a "global model"). This approach would seem to be simple and avoid the complications of model selection. The first serious drawback here is the lack of precision in the estimated parameters. A given data set has only a finite amount of information; each time a parameter estimate is made, the information left is reduced. Increasing the number of parameters eventually makes the fitted model unstable and uninformative. The probability of finding effects (factors) that are actually spurious increases. New parameters are estimated but with increasing uncertainty—this phenomenon is an aspect of the Principle of Parsimony and is closely related to the age-old notion of Occam's razor.

A second serious drawback arises when, as is common, the global model has many parameters in it (dozens, sometimes hundreds). One has to resort to analyzing the set of resultant parameter estimates, as if they were now the data, in order to understand the results of fitting the global model. In essence, one is then fitting reduced-dimension models to the set of (poorly estimated and correlated) global parameter estimates. It is very demanding to do this efficiently and validly. Indeed, the proper way to proceed is to fit the corresponding reduced models (as special cases of the global model) to the original data and do proper multimodel inference. This latter approach facilitates understanding of the information in the data; fitting only a large global model generally fails as a strategy for effective inference.

11. *None of the hypotheses have merit*. AICc ranks the models in the set of alternatives; if none have merit, the models are still ranked. Thus, one needs some measure of the "worth" of either the global model or the model estimated to be best. Thus, standard statistical methods are needed to gauge this matter; these include adjusted R^2, goodness-of-fit tests, and the analysis of regression residuals. We have seen examples where the P value was 0.002 ("highly significant") but the R^2 value was only 0.06. Clearly, if only 6% of the variation in the response variable was in common with the variation in the predictor variables, then little has been learned, even if the P value was "highly significant." Hard thinking in defining the alternative hypotheses is a guard against the case where none of the hypotheses/models are of any inferential value.
12. *Relevant hypotheses and over use of P values*. A review of the papers published in *Ecology* and the *Journal of Wildlife Management* indicated a serious overuse and misuse of null hypothesis testing (Anderson et al. 2000). Many authors discuss P values as if they are evidential; they are not (see Royall 1997). A number of papers reported on hundreds of null hypothesis tests (e.g., as many as 408 P values in a single paper). Hundreds of the null hypotheses were trivial and surely could be rejected on simple *a priori* grounds. The key issue was the failure to explore more relevant questions and to report more informative summary statistics such as the estimated effect size and measures of its precision and evidence ratios.
13. *Contingency Tables*. Integer data (e.g., counts) are often summarized as a contingency table and analyzed using procedures that result in test statistics that are asymptotically χ^2 distributed. If the counts are at all large, these tests are very powerful in rejecting the null. However, one should not jump too fast in claiming that these results indicate that something important has been found. It is important to remember that the alternative hypothesis is never tested. A related issue here is the fact that many sets comprised of count data are overdispersed. This is a bigger issue than we can address here; however, Burnham and Anderson (2002) provide a summary with references to the primary literature.
14. *AICc suggests this model fits the data best.* This statement is incorrect as models (even those within the set) with still more parameters will often fit the data better still. The concept of parsimony enters here and AICc is suggesting that a particular model is best in the sense of trading-off bias versus variance of the fitted model parameters, for a given sample size (i.e., "best" in the K-L information sense).
15. *Debate concerning model likelihoods and model probabilities*. There have been questions raised concerning model likelihoods and model probabilities in the I-T approaches. The claim has been that these quantities are somehow "informal" or not based on sound theory. Interestingly, some of the path to understanding these issues comes from Bayesian results.

The concept of a likelihood for a fitted model seems compelling. That is, the field of statistics ought to be able to extend the useful idea of a "likelihood" (as data-based evidence about something unknown) to models. However, a likelihood is not a derived (as in a result of a theorem) result. It seems Fisher intuited it, then proved likelihood-based inference had good properties (e.g., second order efficiency and consistency). Extending the concept to a model also seems to be a matter of finding (intuiting) a reasonable result. However, using a Bayesian framework can be quite helpful here.

Akaike explored this issue (and model probabilities, to a limited extent) in several papers, most notably Akaike (1979, 1985). Akaike (1979, p. 239) states, "… it is natural to consider exp(-(1/2)AIC) as the "likelihood" of the model determined by the method of maximum likelihood." He continues on page 242, "The numerical results reported in this paper suggest that exp(-(1/2)AIC) plays almost exactly the role of the likelihood expected in a Bayesian procedure." Akaike (1978a, pp. 299–301) provides a Bayesian result regarding the posterior model probabilities.

Also relevant is Akaike (1978b, p. 14), "If the choice of one single model is not the sole purpose of the analysis of the data the average of the models with respect to the approximate posterior probability Cexp {(−1/2)AIC(k)} will provide a better estimate of the true distribution of Y." In other words, the w_i are probabilities. A suitable choice of priors on models

would remove the "approximate" aspect (Burnham and Anderson 2004, pp. 302–305).

If one accepts BIC as a basis for a large-sample approximation to the Bayes factor, then a simple justification of these weights as probabilities, and as the model likelihood, is given in Burnham and Anderson (2002, pp. 302–305) and in Burnham and Anderson (2004, pp. 280–281).

We affirm the meaning of these probabilities and say the Akaike "weights" are probabilities because each is within the interval [0, 1] and they sum to 1. The key issue should be, what do they mean as probabilities? That issue applies as well to the Bayesian framework. Therein prior probabilities on a discrete set of models are probabilities simply by being bounded [0, 1] and summing to 1. We maintain we can easily know if a finite set of numbers is a discrete probability distribution; it is not required that we justify this by how the numbers were "derived" (witness most Bayesian priors). While it is very useful to know these weights can be justified as posterior probabilities, the real issue is being able to say what they mean for model selection and inference (see technical issue 6, above).

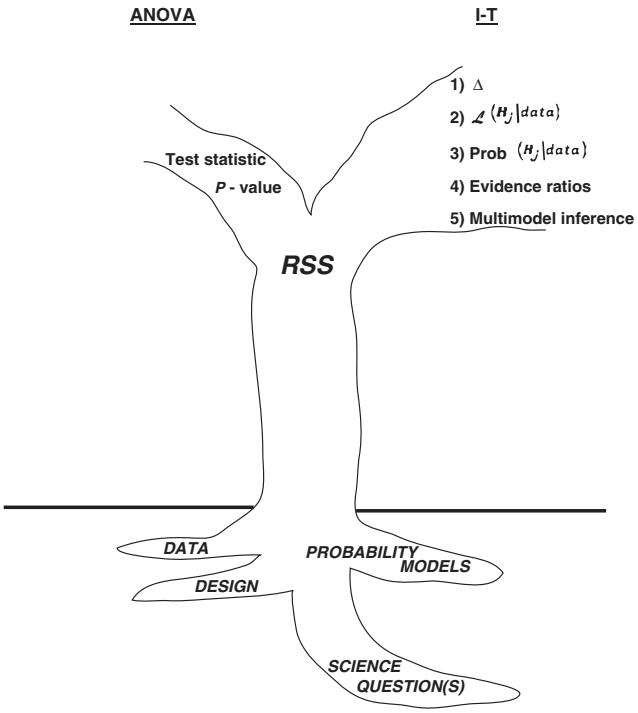

Fig. 3 I-T approaches provide a superior alternative to the traditional test statistic and *P* value paradigm (e.g., *t* tests and ANOVA tables). A conceptual diagram of the pivotal branching point in the tree is the RSS. All the important issues that precede data analysis (the roots) are the same under either analysis approach, including the estimated effect size and its precision for a given model

A simple alternative to ANOVA tables and *P* values

The I-T approaches can also be used for problems traditionally analyzed by analysis of variance (ANOVA) tables with the traditional test statistic, its asymptotic distribution, its *P* value, and an arbitrary judgment as to its *significance*. During the computations leading to a *t* test or ANOVA table one calculates a residual sum of squares (RSS) and this is a branching point that leads to the new information-theoretic approaches and all their advantages in terms of evidence (Fig. 3). Starting with the RSS, the procedure is quite straightforward.

Consider an experiment with random assignment of experimental units to treatment classes, replication, and a proper design involving treatment and control groups (e.g., completely randomized, randomized complete block, factorial). We will focus on a completely randomized design as an example. Data are collected and the traditional analysis involves an ANOVA table. This table summarizes sums of squares for treatment, error (or residual), and total. Division by appropriate degrees of freedom provides mean squares and ratios of these lead to an *F* value and a *P* value. For illustration, let the test statistic (*F*) be 8.1 with appropriate degrees of freedom, the *P* value is 0.009, and this is deemed "statistically significant." The *P* value, the probability of the test statistic being 8.1 or greater, is 0.009, *given the null*. Because the probability is low, one concludes, by default, that the alternative is a better choice (i.e., "significant"). However, the alternative is never tested and the probability of the null and the probability of the alternative are not known.

Information-theoretic approaches provide an attractive alternative to the traditional presentation of *t* tests, ANOVAs and multiple comparisons (means separation tests). Let everything above be the same except that two models are examined: one without a treatment effect and the other with a treatment effect. The first model has two parameters in this case: βo and σ^2 while the second model has 3 parameters: βo, β_1 and σ^2, where β_1 represents the unknown effect size. The essential quantity for each model that is needed for the I-T approach the residual sum of squares, RSS (sometimes called the error sum of squares) as one can then compute the other quantities needed,

$$\mathrm{AICc} = n\log(\mathrm{RSS}/n) + 2K + (2K(K+1))/(n-K-1),$$

where K=2 and 3, respectively. Then, $\Delta_i = \mathrm{AICc}_i - \mathrm{AICc}_{\min}$ and the two model probabilities can be easily computed. Thus, the analyst has the probability of both the null and alternative hypotheses, *given the data*. The strength of evidence can also be gauged by an evidence ratio. In this case, the results are conditioned on the data and all the models of interest; that is why data are collected. Conditioning only on the null model,

which is often no more than a straw man, is an inefficient approach to learning from the data.

A one-way ANOVA is equivalent to an unpaired *t* test and a two-way ANOVA is (or can be) equivalent to a paired *t* test. There are other equivalencies that arise. A detailed example of the paired *t* test case is available from the authors (also see www.Springer.com/978-0-387-74073-7 for exercise 8 in chapter 4). Use of AICc and model selection ideas in an ANOVA framework, rather than classical multiple comparisons methods, is considered by Dayton (1988), and was first suggested by Sugiura (1978).

In summary, both approaches share all the preliminary issues up to the residual (or "error") sums of squares (RSS). Then the traditional and information-theoretic approaches diverge (Fig. 3). In both cases, the computations are simple; however, the inferential information is quite different. Of course, both procedures provide the same estimate of the effect size and its precision for any given model. Given a choice, it seems one would always prefer having the model probabilities (that is, Prob$\{H_o|\text{data}\}$ and Prob$\{H_a|\text{data}\}$) and the evidence ratio rather than just a *P* value which is conditioned on the null.

Problems with null hypothesis testing approaches

Many null hypotheses are trivial (so-called silly nulls) and uninteresting and reflect a lack of thinking about *plausible* alternatives. Finding little/no support for the null does little to provide evidence for the alternative (e.g., perhaps the alternative does not fit either!). At best, the *P* value is not a proper strength of evidence (Royall 1997). The meaning of a *P* value is inferentially odd, Prob{data|null}, and we find that people want to twist this to pretend that a *P* value is Prob{null|data}.

Null hypothesis testing should not be used for observational studies as the distribution of the test statistics under the null is not known (and cannot be gotten by bootstrapping or various Monte Carlo techniques). These traditional methods are especially poor for model building (e.g., step-up, step-down, and stepwise procedures) due to the multiple testing problem.

An array of technical problems arise with the null hypothesis testing approach. One example is the performance of the test when sample size goes to infinity (asymptotic). Consider a simple *t* test where the treatment effect is exactly 0. Then even with infinitely large sample sizes, the procedure will still error at the α-level (e.g., 5% of the time). Moreover, a reported *P* value, or the rejection (or not) of the null hypothesis is uninformative about the actual effect size, at any sample size. Additional information can be found at the websites warnercnr.colostate.edu/~anderson/thompson1.html and warnercnr.colostate.edu/~anderson/nester.html.

In summary, traditional testing approaches leave an analyst without ways to rank hypotheses, cope with data from observational studies, cope with non-nested models, average models, estimate model selection uncertainty, incorporate model selection uncertainty into estimates of precision, or provide confidence sets on models. The I-T methods offer a fundamentally sound, intuitively appealing approach to analysis. Most practicing statisticians have "moved on" to either I-T methods or Bayesian methods; we recommend that people in the life sciences continue to learn about and to adapt the new approaches in their work.

Final thoughts

We offer a few speculations regarding the future, based on experience in the past. One might think that Akaike's main legacy would be his AIC and the related advances. However, Akaike has said that his main contribution has been to point out the importance of defining alternative hypotheses and the related modeling. It is this step that lays the foundations for good science. We need to better develop a culture of hard thinking. Journal editors can play an important role here.

Akaike's later works are more readable for nonstatisticians and more philosophical (Akaike 1983a, b, 1985, 1992, 1994) and draw additional insights from the concept of entropy. Parzen (1994) and Findley and Parzen (1995) provide more information on Akaike and his collected works are listed in Parzen et al. (1998).

The information-theoretic approaches are far more than data analysis. The new methods represent a package, starting with the careful delineation of a worthy science question and ending with the ability to both quantify and qualify the evidence for the set of alternative hypotheses.

Modeling of the alternative hypotheses remains a potential stumbling block as relatively few students in the life sciences have this background in their education. Better education is needed in quantitative methods, including subjects such as calculus, matrix algebra, probability, and mathematical statistics. In addition, education is failing in science history and philosophy: people can receive a doctor of philosophy degree without ever taking a course in science philosophy.

The key ingredients needed for the I-T approaches are the residual sum of squares (RSS) in a least squares framework or the maximized log-likelihood in a likelihood-based analysis. However, further advances include generalized estimating equations (see Qin and Lawless 1994 and Pan 2001a and b). In addition Rissanen (2007) provides alternative approaches also based on information theory,

but from a coding theory standpoint, whereas Konishi and Kitagawa (2007) continue to emphasize the link between information theory and statistical modeling.

There is no excuse for gathering poor data in the twenty-first century. Proper sampling and experimental design are the subjects of hundreds of books and thousands of journal papers. While many people think of "statistics" as only data analysis (or worse yet, just null hypothesis testing and its *P* values), consultation with a statistician can often help in the planning and design stages of a research program.

Quantifying and qualifying the evidence is critically important and is simple using the new approaches. Then the *a priori* set evolves by dropping hypotheses judged to be implausible, refining the remaining hypotheses, and adding new hypotheses based on the earlier evidence. This is a science strategy that promotes fast learning and deep understanding.

Acknowledgments The authors thank The Colorado Cooperative Fish and Wildlife Research Unit and the Department of Fish, Wildlife, and Conservation Biology at Colorado State University for continuous support. We appreciate the help of Robert Montgomerie and that of two anonymous reviewers as these helped improve the manuscript.

References

Akaike H (1973) Information theory as an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) Second international symposium on information theory. Akademiai Kiado, Budapest, pp 267–281

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr AC 19:716–723

Akaike H (1977) On entropy maximization principle. In: Krishnaiah PR (ed) Applications of statistics. North-Holland, Amsterdam, pp 27–41

Akaike H (1978a) On the likelihood of a time series model. Statistician 27:217–235

Akaike H (1978b) Bayesian analysis of the minimum AIC procedure. Ann Inst Stat Math, Part A 30:9–14

Akaike H (1979) A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. Biometrika 66:237–242

Akaike H (1981a) Likelihood of a model and information criteria. J Econom 16:3–14

Akaike H (1981b) Modern development of statistical methods. In: Eykhoff P (ed) Trends and progress in system identification. Pergamon, Paris, pp 169–184

Akaike H (1983a) Statistical inference and measurement of entropy. In: Box GEP, Leonard T, Wu C-F (eds) Scientific inference, data analysis, and robustness. Academic, London, pp 165–189

Akaike H (1983b) Information measures and model selection. Int Statistical Institute 44:277–291

Akaike H (1985) Prediction and entropy. In: Atkinson AC, Fienberg SE (eds) A celebration of statistics. Springer, New York, pp 1–24

Akaike H (1992) Information theory and an extension of the maximum likelihood principle. In: Kotz S, Johnson NL (eds) Breakthroughs in statistics, vol 1. Springer, London, pp 610–624

Akaike H (1994) Implications of the informational point of view on the development of statistical science. In: Bozdogan H (ed) Engineering and scientific applications. Vol. 3, Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. Kluwer, Dordrecht, pp 27–38

Anderson DR (2008) Model based inference in the life sciences: a primer on evidence. Springer, New York

Anderson DR, Burnham KP, Thompson WL (2000) Null hypothesis testing: problems, prevalence, and an alternative. J Wildl Manage 64:912–923

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

Brylawski AMZ, Whittingham LA (2004) An experimental study of mate guarding and paternity in house wrens. Anim Behav 68:1417–1424

Burnham KP, Anderson DR (1992) Data-based selection of an appropriate biological model: the key to modern data analysis. In: McCullough DR, Barrett RH (eds) Wildlife 2001: populations. Elsevier, London, pp 16–30

Burnham KP, Anderson DR (2001) Kullback-Leibler information as a basis for strong inference in ecological studies. Wildlife Research 28:111–119

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Springer, New York

Burnham KP, Anderson DR (2004) Multimodel inference: understanding AIC and BIC in model selection. Sociol Methods Res 33:261–304

Chamberlin TC (1890) The method of multiple working hypotheses. Science 15:92–96, Reprinted 1965, Science 148:754-759

Cover TM, Thomas JA (2006) Elements of information theory, 2nd edn. Wiley, New York

Currie D, Krupa AP, Burke T, Thompson DBA (1999) The effect of experimental male removals on extrapair paternity in the wheatear, *Oenanthe oenanthe*. Anim Behav 57:145–152

Dayton CM (1988) Information criteria for the paired-comparisons problem. Am Stat 52:144–151

Dochtermann N, Jenkins SH (2010) Developing and evaluating candidate hypotheses in behavioral ecology. Behav Ecol Sociobiol doi:10.1007/s00265-010-1039-4

Edwards AWF (1992) Likelihood: expanded edition. Johns Hopkins University Press, Baltimore

Eimes JA, Parker PG, Brown JL, Brown ER (2005) Extrapair fertilization and genetic similarity of social mates in the Mexican jay. Behav Ecol 16:456–460

Elliott LP, Brook BW (2007) Revisiting Chamberlin (1890): multiple working hypotheses for the 21st century. Bioscience 57:608–614

Findley DF, Parzen E (1995) A conversation with Hirotugu Akaike. Stat Sci 10:104–117

Freckleton RP (2010) Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. Behav Ecol Sociobiol doi:10.1007/s00265-010-1045-6

Garamszegi LZ (2010) Information-theoretic approaches to statistical analysis in behavioral ecology: an introduction. Behav Ecol Sociobiol doi:10.1007/s00265-010-1028-7

Griffith SC, Owens IPF, Thuman KA (2002) Extra-pair paternity in birds: a review of interspecific variation and adaptive function. Mol Ecol 11:2195–2212

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

Hall ML (2004) A review of hypotheses for the functions of avian duetting. Behav Ecol Sociobiol 55:415–430

Hobson A, Cheng B-K (1973) A comparison of the Shannon and Kullback information measures. J Stat Phys 7:301–310

Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial (with discussion). Stat Sci 14:382–417

Hoi-Leitner M, Hoi H, Romero-Pujante M, Valera F (1999) Female extra-pair behaviour and environmental quality in the serin (*Serinus serinus*): a test of the 'Constrained Female Hypothesis'. Proc Biol Sci 266:1021–1026

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

Jessop A (1995) Informed assessments: an introduction to information, entropy and statistics. Ellis Horwood, London

Konishi S, Kitagawa G (2007) Information criteria and statistical modeling. Springer, New York

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86

Lack D (1968) Ecological adaptations for breeding in birds. Methuen and Company, London

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall/CRC, Boca Raton

Mennill DJ, Ramsay SM, Boag PT, Ratcliffe LM (2004) Patterns of extrapair mating in relation to male dominance status and female nest placement in black-capped chickadees. Behav Ecol 15:757–765

Mundry R, Numm CL (2009) Stepwise model fitting and statistical inference: turning noise into signal pollution. Am Nat 173:119–123

Neudorf DLH (2004) Extrapair paternity in birds: understanding variation among species. Auk 121:302–307

Pan W (2001a) Akaike's information criterion in generalized estimating equations. Biometrics 57:120–125

Pan W (2001b) Model selection in estimating equations. Biometrics 57:529–534

Parzen E (1994) Hirotugu Akaike, statistical scientist. In: Bozdogan H (ed) Engineering and Scientific Applications, vol. 1. Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach. Kluwer, Dordrecht, Netherlands. pp. 25–32

Parzen E, Tanabe K, Kitagawa G (1998) Selected papers of Hirotugu Akaike. Springer, New York

Pawitan Y (2001) In all likelihood: statistical modelling and inference using likelihood. Oxford Science, Oxford

Perreault S, Lemon RE, Kuhnlein U (1997) Patterns and correlates of extrapair paternity in American redstarts (*Setophaga ruticilla*). Behav Ecol 8:612–621

Platt JR (1964) Strong inference. Science 146:347–353

Qin J, Lawless G (1994) Empirical likelihood and general estimating equations. Ann Stat 22:300–325

Richardson DS, Burke T (2001) Extrapair paternity and variance in reproductive success related to breeding density in Bullock's orioles. Anim Behav 62:519–525

Rissanen J (2007) Information and complexity in statistical modeling. Springer, New York

Royall RM (1997) Statistical evidence: a likelihood paradigm. Chapman and Hall, London

Rubinstein DR (2007) Territory quality drives intraspecific patterns of extrapair paternity. Behav Ecol 18:1058–1064

Schmoll T, Quellmalz A, Dietrich V, Winkel W, Epplen JT, Lubjuhn T (2005) Genetic similarity between pair mates is not related to extrapair paternity in the socially monogamous coal tit. Anim Behav 69:1013–1022

Schmoll T, Mund V, Dietrich-Bischoff V, Winkel W, Lubjuhn T (2007) Male age predicts extrapair and total fertilization success in the socially monogamous coal tit. Behav Ecol 18:1073–1081

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Sellke T, Bayarri MJ, Berger JO (2001) Calibration of p values for testing precise null hypotheses. Am Stat 55:62–71

Severini TA (2000) Likelihood methods in statistics. Oxford University Press, Oxford

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423, and 27:623–656

Soofi ES (1994) Capturing the intangible concept of information. J Am Stat Assoc 89:1243–1254

Steidl RJ (2007) Model selection, hypothesis testing, and risks of condemning analytical tools. J Wildl Manage 70:1497–1498

Sugiura N (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. Commun Stat, Theory Methods A7:13–26

Symonds M, Moussalli A (2010) Model selection, multimodel inference and model averaging using Akaike's information criterion: an introduction for statistically terrified behavioural ecologists. Behavioral Ecology and Sociobiology doi:10.1007/s00265-010-1037-6

Westneat DF, Stewart IRK (2003) Extra-pair paternity in birds: causes, correlates and conflict. Ann Rev Ecolog Syst 34:365–396

Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP (2006) Why do we still use stepwise modeling in ecology and behaviour? J Anim Ecol 75:1182–1189

Yezerinac SM, Weatherhead PJ (1997) Extra-pair mating, male plumage coloration and sexual selection in yellow warblers (*Dendroica petechia*). Proc R Soc Lond B 264:527–532