Preface

We wrote this book to introduce graduate students and research workers in various scientific disciplines to the use of information-theoretic approaches in the analysis of empirical data. These methods allow the data-based selection of a "best" model and a ranking and weighting of the remaining models in a pre-defined set. Traditional statistical inference can then be based on this selected best model. However, we now emphasize that information-theoretic approaches allow formal inference to be based on more than one model (multimodel inference). Such procedures lead to more robust inferences in many cases, and we advocate these approaches throughout the book.

The second edition was prepared with three goals in mind. First, we have tried to improve the presentation of the material. Boxes now highlight essential expressions and points. Some reorganization has been done to improve the flow of concepts, and a new chapter has been added. Chapters 2 and 4 have been streamlined in view of the detailed theory provided in Chapter 7. Second, concepts related to making formal inferences from more than one model (multimodel inference) have been emphasized throughout the book, but particularly in Chapters 4, 5, and 6. Third, new technical material has been added to Chapters 5 and 6. Well over 100 new references to the technical literature are given. These changes result primarily from our experiences while giving several seminars, workshops, and graduate courses on material in the first edition. In addition, we have done substantially more thinking about the issue and reading the literature since writing the first edition, and these activities have led to further insights.

Information theory includes the celebrated Kullback–Leibler "distance" between two models (actually, probability distributions), and this represents a

fundamental quantity in science. In 1973, Hirotugu Akaike derived an estimator of the (relative) expectation of Kullback–Leibler distance based on Fisher's maximized log-likelihood. His measure, now called *Akaike's information criterion* (AIC), provided a new paradigm for model selection in the analysis of empirical data. His approach, with a fundamental link to information theory, is relatively simple and easy to use in practice, but little taught in statistics classes and far less understood in the applied sciences than should be the case.

We do not accept the notion that there is a simple "true model" in the biological sciences. Instead, we view modeling as an exercise in the approximation of the explainable information in the empirical data, in the context of the data being a sample from some well-defined population or process. Rexstad (2001) views modeling as a fabric in the tapestry of science. Selection of a best approximating model represents the inference from the data and tells us what "effects" (represented by parameters) can be supported by the data. We focus on Akaike's information criterion (and various extensions) for selection of a parsimonious model as a basis for statistical inference. Model selection based on information theory represents a quite different approach in the statistical sciences, and the resulting selected model may differ substantially from model selection based on some form of statistical null hypothesis testing.

We recommend the information-theoretic approach for the analysis of data from observational studies. In this broad class of studies, we find that all the various hypothesis-testing approaches have no theoretical justification and may often perform poorly. For classic experiments (control–treatment, with randomization and replication) we generally support the traditional approaches (e.g., analysis of variance); there is a very large literature on this classic subject. However, for complex experiments we suggest consideration of fitting explanatory models, hence on estimation of the size and precision of the treatment effects and on parsimony, with far less emphasis on "tests" of null hypotheses, leading to the arbitrary classification "significant" versus "not significant." Instead, a strength of evidence approach is advocated.

We do not claim that the information-theoretic methods are always the very best for a particular situation. They do represent a unified and rigorous theory, an extension of likelihood theory, an important application of information theory, and they are objective and practical to employ across a very wide class of empirical problems. Inference from multiple models, or the selection of a single "best" model, by methods based on the Kullback–Leibler distance are almost certainly better than other methods commonly in use now (e.g., null hypothesis testing of various sorts, the use of R^2, or merely the use of just one available model). In particular, subjective data dredging leads to overfitted models and the attendant problems in inference, and is to be strongly discouraged, at least in more confirmatory studies.

Parameter estimation has been viewed as an optimization problem for at least eight decades (e.g., maximize the log-likelihood or minimize the residual sum of squared deviations). Akaike viewed his AIC and model selection as "... a natural extension of the classical maximum likelihood principle." This

extension brings model selection and parameter estimation under a common framework—optimization. However, the paradigm described in this book goes beyond merely the computation and interpretation of AIC to select a parsimonious model for inference from empirical data; it refocuses increased attention on a variety of considerations and modeling prior to the actual analysis of data. Model selection, under the information-theoretic approach presented here, attempts to identify the (likely) best model, orders the models from best to worst, and produces a weight of evidence that each model is really the best as an inference.

Several methods are given that allow model selection uncertainty to be incorporated into estimates of precision (i.e., multimodel inference). Our intention is to present and illustrate a consistent methodology that treats model formulation, model selection, estimation of model parameters and their uncertainty in a unified manner, under a compelling common framework. We review and explain other information criteria (e.g., AIC_c, $QAIC_c$, and TIC) and present several examples to illustrate various technical issues, including some comparisons with BIC, a type of dimension consistent criterion. In addition, we provide many references to the technical literature for those wishing to read further on these topics.

This is an applied book written primarily for biologists and statisticians using models for making inferences from empirical data. This is primarily a science book; we say relatively little about decision making in management or management science. Research biologists working either in the field or in the laboratory will find simple methods that are likely to be useful in their investigations. Researchers in other life sciences, econometrics, the social sciences, and medicine might also find the material useful but will have to deal with examples that have been taken largely from ecological studies of free-ranging vertebrates, as these are our interests. Applied statisticians might consider the information-theoretic methods presented here quite useful and a superior alternative to the null hypothesis testing approach that has become so tortuous and uninformative. We hope material such as this will find its way into classrooms where applied data analysis and associated science philosophy are taught. This book might be useful as a text for a course for students with substantial experience and education in statistics and applied data analysis. A second primary audience includes honors or graduate students in the biological, medical, or statistical sciences. Those interested in the empirical sciences will find this material useful because it offers an effective alternative to (1) the widely taught, yet often both complex and uninformative, null hypothesis testing approaches and (2) the far less taught, but potentially very useful, Bayesian approaches.

Readers should ideally have some maturity in the quantitative sciences and experience in data analysis. Several courses in contemporary statistical theory and methods as well as some philosophy of science would be particularly useful in understanding the material. Some exposure to likelihood theory is nearly essential, but those with experience only in least squares regression modeling will gain some useful insights. Biologists working in a team situation with

someone in the quantitative sciences might also find the material to be useful. The book is meant to be relatively easy to read and understand, but the conceptual issues may preclude beginners. Chapters 1–4 are recommended for all readers because they provide the essential material, including concepts of multimodel inference. Chapters 5 and 6 present more difficult material and some new research results. Few readers will be able to absorb the concepts presented here after just one reading of the material; some rereading and additional consideration will often be necessary to understand the deeper points. Underlying theory is presented in Chapter 7, and this material is much deeper and more mathematical. A high-level summary of the main points of the book is provided in Chapter 8.

We intend to remain active in this subject area after this second edition has been published, and we invite comments from colleagues as an ideal way to learn more and understand differing points of view. We hope that the text does not appear too dogmatic or idealized. We have tried to synthesize concepts that we believe are important and incorporate these as recommendations or advice in several of the chapters. This book is an effort to explore the K-L–based multimodel inference in some depth. We realize that there are other approaches, and that some people may still wish to test null hypotheses as the basis for building models of empirical data, and that others may have a more lenient attitude toward data dredging than we advocate here. We do not want to deny other model selection methods, such as cross-validation, nor deny the value of Bayesian methods. Indeed, we just learned (March, 2002) that AIC can be derived as a Bayesian result and have added a note on this issue while reviewing the final page proofs (see Section 6.4.5). However, in the context of objective science, we are compelled by the a priori approach of building candidate models to represent research hypotheses, the use of information-theoretic criteria as a basis for selecting a best approximating model; model averaging, or other multimodel inference methods, when truth is surely very complex; the use of likelihood theory for deriving parameter estimators; and incorporating model selection uncertainty into statistical inferences. In particular, we recommend moving beyond mere selection of a single best model by using concepts and methods of multimodel inference.

Several people have helped us as we prepared the two editions of this book. In particular, we acknowledge C. Chatfield, C. Hurvich, B. Morgan, D. Otis, J. Rotella, R. Shibata, and K. Wilson for comments on earlier drafts of the original manuscript. We are grateful to three anonymous reviewers for comments that allowed us to improve the first edition. D. Otis and W. Thompson served as the reviewers for the second edition and offered many suggestions that were helpful; we greatly appreciate their excellent suggestions. Early discussions with S. Buckland, R. Davis, R. Shibata, and G. White were very useful. S. Beck, K. Bestgen, D. Beyers, L. Ellison, A. Franklin, W. Gasaway, B. Lubow, C. McCarty, M. Miller, and T. Shenk provided comments and insights as part of a graduate course on model selection methods that they took from the authors. C. Flather allowed us to use his data on species accumu-

lation curves as our first example, and we thank C. Braun and the Colorado Division of Wildlife for the data on sage grouse; these data were analyzed by M. Zablan under the supervision of G. White. C. Southwell allowed us to use his kangaroo data from Wallaby Creek. P. Lukacs conducted the bootstrap analysis and some of the Monte Carlo studies of the body fat data in Chapter 5. J. Kullback allowed us to use a photo of his father, and H. Akaike, R. Leibler, R. Shibata, and K. Takeuchi kindly sent us photos and biographical material that appear in the book. Chelsea Publishing Company allowed our use of the photo of L. Boltzmann from the book *Wissenschaftliche Abhandlungen von Ludwig Boltzmann*, and the International Biometric Society authorized our use of a photo of R. Fisher (from *Biometrics* 1964, taken in 1946 by A. Norton). J. Barandun provided the toad photos for the cover, K. Allred provided the cover design, and B. Schmidt helped in coordination. C. Dion, R. Fulton, S. Kane, B. Klein, A. Lyman, and T. Sundlov helped obtain library materials. J. Kimmel and L. Farkas helped in countless ways as we prepared both editions of this book.

We are happy to acknowledge the long-term cooperators of the Colorado Cooperative Fish and Wildlife Research Unit: the Colorado Division of Wildlife, Colorado State University, the Biological Resources Division of the U.S. Geological Survey, and the Wildlife Management Institute. Graduate students and faculty within the Department of Fisheries and Wildlife Biology at Colorado State University provided a forum for our interests in the analysis of empirical data. We extend our appreciation to several federal agencies within the Department of the Interior, particularly the U.S. Geological Survey, for their support of our long-term research interests.

Fort Collins, Colorado

Kenneth P. Burnham
David R. Anderson
January 2002

Contents

8

Summary

This book covers some philosophy about data analysis, some theory at the interface between mathematical statistics and information theory, and some practical statistical methodology useful in the applied sciences. In particular, we present a general strategy for modeling and data analysis. We provide some challenging examples from our fields of interest, provide our ideas as to what not to do, and suggest some areas needing further theoretical development. We side with the fast-growing ranks that see limited utility in statistical null hypothesis testing. Finally, we provide references from the diverse literature on these subjects for those wishing to study further.

Conceptually, there is information in the observed data, and we want to express this information in a compact form via a "model." Such a model represents a scientific hypothesis and is then a basis for making inferences about the process or system that generated the data. One can view modeling of information in data as a change in "coding" like a change in language. A concept or emotion expressed in one language (e.g., French) loses some exactness when expressed in another language (e.g., Russian). A given set of data has only a finite, fixed amount of information. The (unachievable) goal of model selection is to attain a perfect 1-to-1 translation such that no information is lost in going from the data to a model of the information in the data. Models are only approximations, and we cannot hope to perfectly achieve this idealized goal. However, we can attempt to find a model of the data that is best in the sense that the model loses as little information as possible. This thinking leads directly to Kullback–Leibler information $I(f, g)$: the information lost when model g is used to approximate full reality f. We wish then to select a model that minimizes K-L information loss. Because we must estimate model

parameters from the data, the best we can do is to minimize (estimated) expected K-L information loss. However, this can easily be done using one of the information-theoretic criteria (e.g., AIC, AIC_c, QAIC, or TIC). Then a *good* model allows the efficient and objective separation or filtration of *information* from *noise*. In an important sense, we are not really trying to model the *data*; instead, we are trying to model the *information* in the data.

While we use the notation f to represent truth or full reality, we deny the existence of a "true model" in the life sciences. Conceptually, let f be the process (truth) that generates the sample data we collect. We want to make inferences about truth, while realizing that full reality will always be beyond us when we have only sample data. Data analysis should not be thought of as an attempt to identify f; instead, we must seek models that are good approximations to truth and from which therefore we can make valid inferences concerning truth. We do not want merely to describe the data using a model that has a very large number of parameters; instead, we want to use the data to aid in the selection of a parsimonious model that allows valid inferences to be made about the system or process under study. A parsimonious model, representing a well-defended scientific hypothesis, aids in our understanding of the system of interest.

Relatively few statistics books provide a summary of the key points made and yet fewer provide an effective, unified strategy for data analysis and inference where there is substantial complexity. The breadth of the technical subjects covered here makes a summary difficult to write. Undergraduate students occasionally ask the professor, "What is important for me to know for the final examination?" The professor is typically irritated by such a question. Surely, the student should realize that it is *all* important! Indeed, our interpretation of Akaike's pioneering work is that it *is* all important. The information-theoretic paradigm is a *package*; each of the package's contents is important in itself, but it is the integration of the contents that makes for an effective philosophy, a consistent strategy, and a practical and powerful methodology. The part of this package that has been so frequently left out is the critical thinking, hypothesis generation, and modeling *before* examination of the data; ideally, much of this thinking should occur prior even to data collection. This is the point where the science of the issue formally enters the overall "analysis" (Anderson and Burnham 1999a).

The information-theoretic methods we present can be used to select a single best model that can be used in making inferences from empirical data. AIC is often portrayed in the literature in this simple manner. The general approach is much richer than this simplistic portrayal of model selection might suggest. **In fact, an emphasis of this second edition is multimodel inference (MMI). MMI has several advantages; all relate to the broad subject of model selection uncertainty**. One can easily rank alternative models (hypotheses) from best to worst using the convenient differences Δ_i. The likelihood for each model, given the data [i.e., $\mathcal{L}(g_i \mid data)$], can be easily computed, and these

can be normalized to obtain Akaike weights (w_i) which can be interpreted as probabilities. Confidence sets of models can be defined to aid in identifying a subset of good models. Evidence ratios are useful for comparing relative support of one model versus another, given the data; such ratios are useful, irrespective of other models in the set.

Model selection uncertainty can be easily quantified using Akaike weights (the bootstrap is an alternative). Estimates of this component of uncertainty can be incorporated into unconditional estimates of precision using several methods. For many problems (e.g., prediction) model-averaging has advantages, and we treat this important issue in Chapters 4–5. Thus, we often recommend formal inference from all models in the set.

For those who have scanned through the pages of this book there might be surprise at the general lack of mathematics and formulas (Chapters 6 and 7 being the exceptions). That has been our intent. The *application* of the information-theoretic methods is relatively simple. They are easy to understand and use ("low tech"), while the underlying theory is quite deep (e.g., Chapter 7). As we wrote the book and tried to understand Akaike's various papers (see Parzen et al. 1998) we found the need to delve into various issues that are generally philosophical. The science of the problem has to be brought into modeling *before* one begins to rummage through the data (data dredging). In some critical respects, applied statistics courses are failing to teach statistics as an integral part of scientific discovery, with little about modeling and model selection methods or their importance, while succeeding (perhaps) in teaching null hypothesis testing methods and data analysis methods based on the assumption that the model is both true and given. Sellke et al. (2001:71) note, "The standard approach in teaching—stressing the formal definition of a p value while warning against its misinterpretation—has simply been an abysmal failure." It seems necessary to greatly reduce the reporting of P-values (Anderson et al. 2001b and d).

8.1 The Scientific Question and the Collection of Data

The formulation of the research question is crucial in investigations into complex systems and processes in the life sciences. A good answer to a poor question is a mistake all too often seen in the published literature and is little better than a poor answer to a poor question. Investigators need to continually readdress the importance and quality of the question to be investigated. Good scientific hypotheses, represented by models, must have a place at the head of the table.

A careful program of data collection must follow from the hypotheses posed. Particular attention should be placed on the variables to be measured and interesting covariates. Observational studies, done well, can show patterns,

associations, and relationships and are confirmatory in the sense that certain issues stem from a priori considerations. More causal inference must usually come from more formal experimentation (i.e., important confounding factors are controlled or balanced, experimental units are randomly assigned to treatment and control groups with adequate replication), but see Anderson et al. (1980), Gail (1996), Beyers (1998), and Glymour (1998) for alternative philosophies. Valid inference must assume that these basic important issues have been carefully planned and conducted. Before one should proceed, two general questions must be answered in the affirmative:

Are the study objectives sound, relevant, and achievable?
Has there been proper attention to study design and laboratory or field protocol?

8.2 Actual Thinking and A Priori Modeling

Fitting models, each representing a scientific hypothesis, to data has been important in many biological, ecological, and medical investigations. Then statistical inferences about the system of interest are made from an interpretable parsimonious model of the observational or experimental data. We expect to see this activity increase as more complicated scientific and management issues are addressed. In particular, a priori modeling becomes increasingly important as several data sets are collected on the same issue by different laboratories or at widely differing field sites over several years.

We recommend much more emphasis on thinking! Leave the computer idle for a while, giving time to think hard about the overall problem. What useful information is contained in the published literature, even on issues only somewhat related to the issue at hand? What nonlinearities and threshold effects might be predicted? What interactions are hypothesized to be important? Can two or more variables be combined to give a more meaningful variable for analysis? Should some variables be dropped from consideration? Discussions should be encouraged with the people in the field or laboratory that were close to the data collection. What parameters might be similar across groups (i.e., data sets)? Model building should be driven by the underlying science of the issue combined with a good understanding of mathematical models. Ideally, this important conceptual phase might take several days or even weeks of effort; this seems far more time than is often spent under current practice.

Biologists generally subscribe to the philosophy of "multiple working hypotheses" (Chamberlain 1890, Platt 1964, Mayr 1997), and these should form the basis for the set of candidate models to be considered formally. Model building can begin during the time that the a priori considerations are being sorted out. Modeling must carefully quantify the science hypotheses of interest. Often it is effective to begin with the global model and work toward some lower-dimensional models. Others may favor a bottom-up approach. The critical matter here is that one arrives, eventually, at a small set of good

candidate models, prior to examination of the empirical data. We advise the inclusion of all models that are reasonably justified prior to data analysis; however, every attempt should be made to keep the number of candidate models small.

Critical Thinking

Our science culture does not do enough to regularly expect and enforce critical thinking. This failure has slowed the scientific discovery process.

We fail to fault the trivial content of the typical ecological hypothesis.

There is a need for more *careful thinking* (than is usually evident) and a *better balance* between scientific hypotheses, data, and analysis theory.

Chamberlin's concept of *multiple working hypotheses*, suggested well over 100 years ago, has a deep level of support among science philosophers. He thought the method led to "certain distinctive habits of mind and had prime value in education." Why has this principle not become the standard, rather than the rare exception, in so many fields of applied science?

Platt (1964) noted that years and decades can be wasted on experiments, unless one thinks carefully in advance about what the most important and conclusive experiments would be.

With the information-theoretic approach, there is no concept of a "null" hypothesis, or a statistical hypothesis test, or an arbitrary α-level, or questionable power, or the multiple testing problem, or the fact that the so-called null hypothesis is nearly always *obviously* false in the first place. Much of the application of statistical hypothesis testing arbitrarily classifies differences into meaningless categories of "significant" and "nonsignificant," and this practice has little to contribute to the advancement of science (Anderson et al. 2000). We recommend that researchers stop using the term "significant," since it is so overused, uninformative, and misleading. The results of model selection based on estimates of expected (relative) Kullback–Leibler information can be very different from the results of some form of statistical hypothesis testing (e.g., the simulated starling data, Section 3.4, or the sage grouse data, Section 3.5).

So, investigators may proceed with inferential or confirmatory data analysis if they feel satisfied that they can objectively address two questions:

Was the set of candidate models derived a priori?
What justifies this set?

The justification should include a rationale for models both included and excluded from the set. A carefully defined set of models is crucial whether information-theoretic methods are used to select the single best model, or the entire set of models is used to reach defensible inferences. If so little is known about the system under study that a large number of models must be included in the candidate set, then the analysis should probably be considered only exploratory (if models are developed as data analysis progresses, it is both exploratory and risky). **One should check the fit or adequacy of the global model using standard methods. If the global model is inadequate (after,**

perhaps, adjusting for overdispersed count data), then more thought should be put into model building and thinking harder about the system under study and the data collected. There is no substitute for good, hard thinking at this point (Platt 1964).

8.3 The Basis for Objective Model Selection

Statistical inference from a data set, *given a model*, is well advanced and supported by a very large amount of theory. Theorists and practitioners are routinely employing this theory, either likelihood or least squares, in the solution of problems in the applied sciences. The most compelling question is, "*what model to use*?" Valid inference must usually be based on a good approximating model, but which one?

Akaike chose the celebrated Kullback–Leibler discrimination information as a basis for model selection. This is a fundamental quantity in the sciences and has earlier roots in Boltzmann's concept of *entropy*, a crowning achievement of nineteenth-century science. The K-L distance between conceptual truth f and model g is defined for continuous functions as the integral

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

where log denotes the natural logarithm and f and g are n-dimensional probability distributions. Kullback and Leibler (1951) developed this quantity from "information theory," thus the notation $I(f, g)$ as it relates to the "information" lost when model g is used to approximate truth f. Of course, we seek an approximating model that loses as little information as possible; this is equivalent to minimizing $I(f, g)$ over the models in the set. Full reality is considered to be fixed. An interpretation equivalent to minimizing $I(f, g)$ is that we seek an approximating model that is the "shortest distance" from truth. Both interpretations seem useful and compelling.

The K-L distance can be written equivalently as

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

The two terms on the right in the above expression are statistical expectations with respect to f (truth). Thus, the K-L distance (above) can be expressed as a difference between two expectations,

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

each with respect to the true distribution f. The first expectation, $\mathrm{E}_f[\log(f(x))]$, is a constant that depends only on the unknown true distribution. Therefore, treating this unknown term as a constant, only a measure of *relative* distance

is possible. Then

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

or

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

Thus, the term $\big(I(f, g) - \text{constant}\big)$ is a *relative* distance between truth f and model g. This provides a deep theoretical basis for model selection if one can compute or estimate $\mathrm{E}_f[\log(g(x \mid \theta))]$.

Akaike (1973, 1974, 1985, 1994) showed that the critical quantity for estimating relative K-L information was

$$\mathbf{E}_y\mathbf{E}_x[\mathbf{log}(\boldsymbol{g}(\boldsymbol{x}|\hat{\boldsymbol{\theta}}(\boldsymbol{y})))],$$

where y and x are independent random samples from the same distribution and both statistical expectations are taken with respect to truth (f). This double expectation, both with respect to truth f, is the target of model selection approaches based on K-L information.

8.4 The Principle of Parsimony

Parsimony is the concept that a model should be as simple as possible with respect to the included variables, model structure, and number of parameters. Parsimony is a desired characteristic of a model used for inference, and it is usually visualized as a suitable tradeoff between squared bias and variance of parameter estimators (Figure 1.3). Parsimony lies between the evils of underfitting and overfitting (Forster and Sober 1994, Forster 1999). Expected K-L information is a fundamental basis for achieving proper parsimony in modeling.

The concept of parsimony has a long history in the sciences. Often this is expressed as "Occam's razor": shave away all that is unnecessary. The quest is to make things "as simple or small as possible." Parsimony in statistics represents a tradeoff between bias and variance as a function of the dimension of the model (K). A good model is a proper balance between underfitting and overfitting, given a particular sample size (n). Most model selection methods are based on the concept of a squared bias versus variance tradeoff. Selection of a model from a set of approximating models must employ the concept of parsimony. These philosophical issues are stressed in this book, but it takes some experience and reconsideration to reach a full understanding of their importance.

8.5 Information Criteria as Estimates of Expected Relative Kullback–Leibler Information

Roots of Theory

As deLeeuw (1992) noted, Akaike found a formal relationship between Boltzmann's entropy and Kullback–Leibler information (dominant paradigms in information and coding theory) and maximum likelihood (the dominant paradigm in statistics).

This finding makes it possible to combine estimation (point and interval estimation) and model selection under a single theoretical framework: optimization.

Akaike's (1973) breakthrough was the finding of an estimator of the expected relative K-L information, based on a bias-corrected maximized log-likelihood value. His estimator was an approximation and, under certain conditions, asymptotically unbiased. He found that

$$\textbf{estimated expected (relative) K-L information} \approx \log(\mathcal{L}(\hat{\theta})) - K,$$

where $\log(\mathcal{L}(\hat{\theta}))$ is the maximized log-likelihood value and K is the number of estimable parameters in the approximating model (this is the bias-correction term). Akaike multiplied through by -2 and provided Akaike's information criterion (AIC)

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

Akaike considered his information-theoretic criterion an extension of Fisher's likelihood theory. Conceptually, the principle of parsimony is enforced by the added "penalty" (i.e., $2K$) while minimizing AIC.

Assuming that a set of a priori candidate models has been carefully defined, then AIC is computed for each of the approximating models in the set, and the model where AIC is minimized is selected as best for the empirical data at hand. This is a simple, compelling concept, based on deep theoretical foundations (i.e., K-L information). Given a focus on a priori issues, modeling the relevant scientific hypotheses, and model selection, *the inference is the selected model*. In a sense, parameter estimates are almost byproducts of the selected model. This inference relates to the estimated best approximation to truth and what information seems to be contained in the data.

Important refinements followed shortly after the pioneering work by Akaike. Most relevant was Takeuchi's (1976) information criterion (termed TIC), which provided an asymptotically unbiased estimate of relative expected K-L information. TIC is little used, since it requires the estimation of $K \times K$ matrices of first and second partial derivatives of the log-likelihood function, and its practical use hinges on the availability of a relatively large sample size. In a sense, AIC can be viewed as a parsimonious version of TIC. A second refinement was motivated by Sugiura's (1978) work, and resulted in a series of papers by Hurvich and Tsai (1989, 1990b, 1991, 1994, 1995a and 1995b, 1996). They

provided a second order approximation, termed AIC_c, to estimated, expected relative K-L information,

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

where n is sample size The final bias-correction term vanishes as n gets large with respect to K (and AIC_c becomes AIC), but the additional term is important if n is not large relative to K (we suggest using AIC_c if $n/K < 40$ or, alternatively, always using AIC_c).

A third extension was a simple modification to AIC and AIC_c for overdispersed count data (Lebreton et al. 1992). A variance inflation factor $\hat{c}$ is computed from the goodness-of-fit statistic, divided by its degrees of freedom, $\hat{c} = \chi^2/\text{ df}$. The value of the maximized log-likelihood function is divided by the estimate of overdispersion to provide a proper estimate of the log-likelihood. These criteria are denoted by QAIC and QAIC_c as they are derived from quasi-likelihood theory (Wedderburn 1974),

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

and

$$\begin{aligned}\mathbf{QAIC}_c &= -[2\log(\mathcal{L}(\hat{\theta}))/\hat{c}] + 2K + \frac{2K(K+1)}{n-K-1} \\ &= \mathbf{QAIC} + \frac{2K(K+1)}{n-K-1}.\end{aligned}$$

When no overdispersion exists, $c = 1$, and the formulas for QAIC and QAIC_c reduce to AIC and AIC_c, respectively. There are other, more sophisticated, ways to account for overdispersion in count data, but this simple method is often quite satisfactory. Methods are given in Chapter 6 to allow different partitions of the data to have partition-specific estimates of overdispersion. Note that the number of estimable parameters (K) must include the number of estimates of c. Thus, if males and females have different degrees of overdispersion and these are to be estimated from the data, then K must include 2 parameters for these estimates.

AIC is often presented in the scientific literature in an ad hoc manner, as if the bias-correction term K (the so-called penalty term) was arbitrary. Worse yet, perhaps, is that AIC is often given without reference to its fundamental link with Kullback–Leibler information. Such shallow presentations miss the point, have had very negative effects, and have misled many into thinking that there is a whole class of selection criteria that are "information-theoretic" (Chapter 6). Criteria such as AIC, AIC_c, QAIC, and TIC are estimates of expected (relative) Kullback–Leibler distance and are useful in the analysis of real data in the "noisy" sciences.

8.6 Ranking Alternative Models

Because only *relative* K-L information can be estimated using one of the information criteria, it is convenient to rescale these values such that the model with the minimum AIC (or AIC_c or TIC) has a value of 0. Thus, information-criterion values can be rescaled as simple differences,

$$\begin{aligned}\Delta_i &= AIC_i - AIC_{min} \\ &= \hat{E}_{\hat{\theta}}[\hat{I}(f, g_i)] - \min \hat{E}_{\hat{\theta}}[\hat{I}(f, g_i)].\end{aligned}$$

While the value of minimum $\hat{E}_{\hat{\theta}}[\hat{I}(f, g_i)]$ is not known (only the relative value), we have an estimate of the size of the increments of information loss for the various models compared to the estimated best model (the model with the minimum $E_{\hat{\theta}}[\hat{I}(f, g_i)]$). The 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. As a rough rule of thumb, models having Δ_i within 1–2 of the best model have substantial support and should receive consideration in making inferences. Models having Δ_i within about 4–7 of the best model 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 fail to explain some substantial structural variation in the data. If the observations are not independent (but are treated as such) or if the sample size is quite small, or it there is a very large number of models, then the simple guidelines above cannot be expected to hold.

There are cases where a model with $\Delta_i > 10$ might still be useful, particularly if the sample size is very large (e.g., see Section 6.8.2). For example, let model A, with year-specific structure on one of the parameters, be the best model in the set ($\Delta_A = 0$) and model B, with less structure on the subset of year-specific parameters, have $\Delta_B = 11.4$. Assume that all models in the candidate set were derived prior to data analysis (i.e., no data dredging). Clearly, model A is able to identify important variation in a parameter across years; this is important. However, in terms of understanding and generality of inference based on the data, it might sometimes be justified to use the simpler model B, because it may seem to "capture" the important fixed effects. Models A and B should both be detailed in any resulting publication, but understanding and interpretation might be enhanced using model B, even though some information in the data would be (intentionally) lost. Such lost information could be partially recovered by, for example, using a random effects approach (see Section 3.5.5) to estimate the mean of the time-effects parameter and the variance of its distribution.

The principle of parsimony provides a philosophical basis for model selection; Kullback–Leibler information provides an objective target based on deep, fundamental theory; and the information criteria (particularly AIC and AIC_c) provide a practical, general methodology for use in data analysis. Objective

model selection and model weighting can be rigorously based on these principles. In practice, one need not assume that any "true model" is contained in the set of candidates (although this is sometimes stated, erroneously, in the technical literature). [We note that several "dimension-consistent criteria" have been published that attempt to provide asymptotically unbiased (i.e., "consistent") estimates of the dimension (K) of the "true model." Such criteria are only estimates of K-L information in a strained way, are based on unrealistic assumption sets, and often perform poorly (even toward their stated objective) unless a very large sample size is available (or where σ^2 is negligibly small, such as in many problems in the physical sciences). We do not recommend these dimension-consistent criteria for the analysis of real data in the life sciences.]

8.7 Scaling Alternative Models

The information-theoretic approach does more than merely estimate which model is best for making inference, given the set of a priori candidate models and the data. The Δ_i allow a ranking of the models from an estimated best to the worst; the larger the Δ_i, the less plausible is model i. **In many cases it is not reasonable to expect to be able to make inferences from a single (best) model; biology is not simple; why should we hope for a simple inference from a single model?** The information-theoretic paradigm provides a basis for examination of alternative models and, where appropriate, making formal inference from more than one model (MMI).

The simple transformation $\exp(-\frac{1}{2}\Delta_i)$ results in the (discrete) likelihood of model i, given the data $\mathcal{L}(g_i|x)$. These are functions in the same sense that $\mathcal{L}(\theta|x, g_i)$ is the likelihood of the parameters θ, given the data (x) and the model (g_i). These likelihoods are very useful; for example, the evidence ratio for model i versus model j is merely

$$\mathcal{L}(g_i|x)/\mathcal{L}(g_j|x).$$

It is convenient to normalize these likelihoods such that they sum to 1, as

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

and interpret these as a weight of evidence. Akaike (e.g., Akaike 1978b, 1979, 1980, and 1981b; also see Kishino 1991 and Buckland et al. 1997) suggested these values, and we have found them to be simple and very useful. The evidence ratio of model i versus model j is then just w_i/w_j; this is identical to the ratio of the likelihood $\mathcal{L}(g_i|x)/\mathcal{L}(g_j|x)$. Drawing on Bayesian ideas we can interpret w_i as the estimated probability that model i is the K-L best model for the data at hand, given the set of models considered (see Section 6.4.5).

An interesting and recent finding is that AIC can be derived under a formal Bayesian framework, and this fact has led to some deeper insights. The breakthrough here was to consider priors on models that are a function of both n and K (we call this class of model priors "savvy," i.e., shrewdly informative); then AIC and AIC_c fall out as a strictly Bayesian result. Indeed, as AIC has a Bayesian derivative, it is compelling to interpret the Akaike weights as posterior model probablilities. While many (objective) Bayesians are comfortable with the use of a defuse or noninformative prior on model parameters (e.g., a uniform prior on a model parameter), use of such defuse priors on models (such as $1/R$) may have poor properties or unintended consequences. That is, some priors on models may be uninformative, but not innocent. In the end, the Bayesian derivation of AIC (or AIC_c) and BIC differ only in their priors on models. However, these criteria are fundamentally different in a variety of substantive ways. In this book we place an emphasis on the derivation of AIC and AIC_c as bias-corrected estimates of Kullback–Leibler information because this seems so much more objective and fundamental.

The w_i are useful as the "weight of evidence" in favor of model i as being the actual K-L best model in the set. The bigger the Δ_i, the smaller the weight and the less plausible is model i as being the best approximating model. Inference is conditional on both the data and the set of a priori models considered.

Alternatively, one could draw B bootstrap samples (B should often be 10,000 rather than 1,000), use the appropriate information criterion to select a best model for each of the B samples, and tally the proportion of samples whereby the ith model was selected. Denote such bootstrap-selection frequencies by $\hat{\pi}_i$. While w_i and $\hat{\pi}_i$ are not estimates of exactly the same entity, they are often closely related and provide information concerning the uncertainty in the best model for use. The Akaike weights are simple to compute, while the bootstrap weights are computer-intensive and not practical to compute in some cases (e.g., the simulated starling experiment, Section 3.4), because thousands of bootstrap repetitions must be drawn and analyzed.

Under the hypothesis-testing approach, nothing can generally be said about ranking or scaling models, particularly if the models were not nested. In linear least squares problems one could turn to adjusted R^2 values for a rough ranking of models, but other kinds of models cannot be scaled using this (relatively very poor) approach (see the analogy in Section 2.5).

8.8 MMI: Inference Based on Model Averaging

Rather than base inferences on a single selected best model from an a priori set of models, we can base our inferences on the entire set by using model-averaging. The key to this inference methodology is the Akaike weights. Thus, if a parameter θ is in common over all models (as θ_i in model g_i), or our goal is prediction, by using the weighted average we are basing point inference on

the entire set of models,

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

or

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

This approach has both practical and philosophical advantages. Where a model-averaged estimator can be used, it appears to have better precision and reduced bias compared to $\hat{\theta}$ from the selected best model.

If one has a large number of closely related models, such as in regression-based variable selection (all-subsets selection), designation of a single best model is unsatisfactory, because that estimated "best" model is highly variable from data set to data set. In this situation model-averaging provides a relatively much more stabilized inference. The concept of inference being tied to all the models can be used to reduce model selection bias effects on regression-coefficient estimates in all-subsets selection. For the regression coefficient associated with predictor x_j we use the estimate $\hat{\bar{\beta}}_j$, which is the estimated regression coefficient β_j averaged over all models in which x_j appears:

$$\hat{\bar{\beta}}_j = \frac{\sum_{i=1}^{R} w_i I_j(g_i) \hat{\beta}_{j,i}}{w_+(j)},$$

$$w_+(j) = \sum_{i=1}^{R} w_i I_j(g_i),$$

where i is for model $i = 1, \ldots, R$, j is for predictor variable j, and

$$I_j(g_i) = \begin{cases} 1 \text{ if predictor } x_j \text{ is in model } g_i, \\ 0 \text{ otherwise.} \end{cases}$$

Conditional on model g_i being selected, model selection has the effect of biasing $\hat{\beta}_{j,i}$ away from zero. Thus a new estimator, denoted by $\tilde{\bar{\beta}}_i$, is suggested:

$$\tilde{\bar{\beta}}_i = w_+(i) \hat{\bar{\beta}}_i.$$

Investigation of this idea, and extensions of it, is an open research area. The point here is that while $\hat{\bar{\beta}}_j$ can be computed ignoring models other than the ones x_j appears in, $\tilde{\bar{\beta}}_i$ does require fitting all R of the a priori models.

8.9 MMI: Model Selection Uncertainty

At first, one might think that one could use an information critrion to select an approximating model that was "close" to truth (remembering the bias versus

variance tradeoff and the principle of parsimony) or that "lost the least information" and then proceed to use this selected model for inference as if it had been specified a priori as the only model considered. Actually, this approach would not be terrible, since at least one would have a reasonable model, selected objectively, based on a valid theory and a priori considerations. This approach would often be superior to much of current practice. Except in the case where the best model has an Akaike weight > 0.9, the problem with considering only this model, and the usual measures of precision *conditional on this selected model*, is that this tends to overestimate precision. Breiman (1992) calls the failure to acknowledge model selection uncertainty a "quiet scandal." [We might suggest that the widespread use of statistical hypothesis testing and blatant data dredging in model selection represent "loud scandals."] In fact, there is a variance component due to model selection uncertainty that should be incorporated into estimates of precision such that these are unconditional (on the selected model). While this is a research area needing further development, several useful methods are suggested in this book, and others will surely appear in the technical literature in the next few years, including additional Bayesian approaches.

The Akaike (w_i) or bootstrap (π_i) weights that are used to rank and scale models can also be used to estimate unconditional precision where interest is in the parameter θ over R models (model g_i, for $i = 1, \ldots, R$),

$$\widehat{\mathbf{var}}(\bar{\hat{\theta}}) = \left[\sum_{i=1}^{R} w_i \sqrt{\widehat{\mathbf{var}}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \bar{\hat{\theta}})^2} \right]^2,$$

$$\widehat{\mathbf{var}}(\bar{\hat{\theta}}) = \left[\sum_{i=1}^{R} \pi_i \sqrt{\widehat{\mathbf{var}}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \bar{\hat{\theta}})^2} \right]^2.$$

These estimators, from Buckland et al. (1997), include a term for the conditional sampling variance, given model g_i (denoted by $\widehat{\mathrm{var}}(\hat{\theta}_i \mid g_i)$ here) and incorporate a variance component for model selection uncertainty $(\hat{\theta}_i - \bar{\hat{\theta}})^2$. These estimators of unconditional variance are also appropriate in cases where one wants a model-averaged estimate of the parameter when θ appears in all models.

Chapter 4 gives some procedures for setting confidence intervals that include model selection uncertainty, and it is noted that achieved confidence-interval coverage is then a useful measure of the utility of methods that integrate model selection uncertainty into inference. Only a limited aspect of model uncertainty can be currently handled. *Given* a set of candidate models and an objective selection method, we can assess selection uncertainty. The uncertainty in defining the set of models cannot be addressed; we lack a theory for this issue. In fact, we lack good, general guidelines for defining the a priori set of models. We expect papers to appear on these scientific and philosophical issues in the future.

8.10 MMI: Relative Importance of Predictor Variables

Inference on the importance of a variable is similarly improved by being based on all the models. If one selects the best model and says that the variables in it are the important ones and the other variables are not important, this is a very naive, unreliable inference. We suggest that the relative importance of variable x_j be measured by the sum of the Akaike weights over all models in which that variable appears:

$$w_+(j) = \sum_{i=1}^{R} w_i I_j(g_i).$$

Thus again, proper inference requires fitting all the models and then using a type of model-averaging. A certain balance in the number of models each with model j, must be achieved. When possible, one should use inference based on all the models, via model-averaging and selection bias adjustments, rather than risk making inference based only on the model estimated to be the best and, often, ignoring other models that are also quite good.

8.11 More on Inferences

Information-theoretic methods do not offer a mechanical, unthinking approach to science. While these methods can certainly be misused, they elicit careful thinking as models are developed to represent the multiple scientific hypotheses that must be the focus of the entire study. A central theme of this book is to call attention to the need to ask better scientific questions in the applied sciences (Platt 1964). Rather than test trivial null hypotheses, it is better to ask deeper questions relating to well-defined alternative hypotheses. For this goal to be achieved, a great deal more hard thinking will be required.

There needs to be increased attention to separating those inferences that rest on a priori considerations from those resulting from some degree of data dredging. White (2000:1097) comments, "Data snooping is a dangerous practice to be avoided, but in fact is endemic."

Essentially no justifiable theory exists to estimate precision (or test hypotheses, for those still so inclined) when data dredging has taken place (the theory (mis)used is for a priori analyses, assuming that the model was the only one fit to the data). A major concern here is the finding of effects and relationships that are actually spurious where inferences are made post hoc (see Lindsey 1999b, Anderson et al. 2001b). This glaring fact is either not understood by practitioners and journal editors or is simply ignored. Two types of data dredging include (1) an iterative approach, in which patterns and differences observed after initial analysis are "chased" by repeatedly building new models with these effects included and (2) analysis of "all possible models." Data dredging is a poor approach to making inferences about the sampled population, and both

types of data dredging are best reserved for more exploratory investigations and are not the subject of this book.

The information-theoretic paradigm avoids statistical null hypothesis testing concepts and focuses on relationships of variables (via selection) and on the estimation of effect size and measures of its precision. This paradigm is primarily in the context of making inferences from a single selected model or making robust inference from many models (e.g., using model-averaging based on Akaike weights). Data analysis is a process of learning what effects are supported by the data and the degree of complexity of the best models in the set. Often, models other than just the estimated best model contain valuable information. Evidence ratios and confidence sets on models help in making inferences on all, or several of the best, models in the set. Information-theoretic approaches should not be used unthinkingly; a good set of candidate models is essential, and this involves professional judgment and representation of the scientific hypotheses into the model set.

When the analysis of data has been completed under an information-theoretic approach, one should gather and report on the totality of the evidence at hand. The primary evidence might be the selected model and its parameter estimates and appropriate measures of precision (including a variance component for model selection uncertainty.) The ranks of each of the R models and the Akaike weights should be reported and interpreted. Model-averaged parameter estimates are often important, particularly for prediction. Evidence ratios, confidence sets on the K-L best model, and a ranking of the relative importance of predictor variables are often useful evidence. When appropriate, quantities such as adjusted R^2 and $\hat{\theta}^2$ should be reported for, at least, the best model. The results from an analysis of residuals for the selected model might also be important to report and interpret. Every effort should be made to fully and objectively report on all the evidence available. If some evidence arose during post hoc activities, this should be clearly stated in published results. Figure 8.1 provides a simplistic graphical representation of the information-theoretic approach. The point of Figure 8.1 is to reinforce some foundational issues (bottom building blocks) and the practical tools and methods (middle row of blocks) that rest on these foundations. If these are used carefully and objectively, one can hope to provide compelling evidence allowing valid inferences. The weakest link seems often to be the left block on the bottom—thinking deeply about the science problem and the alternative hypotheses!

It seems worth noting that K-L information and MMI can be used in certain types of conflict resolution where data exist that are central to the possible resolution of the conflict (Anderson et al. 1999, 2001*c*). Details here would take us too far afield; however, as Hoeting et al. (1999) noted (in a Bayesian context), "Model averaging also allows users to incorporate several competing models in the estimation process; thus model averaging may offer a committee of scientists a better estimation method than the traditional approach of trying to get the committee to agree on a best model."

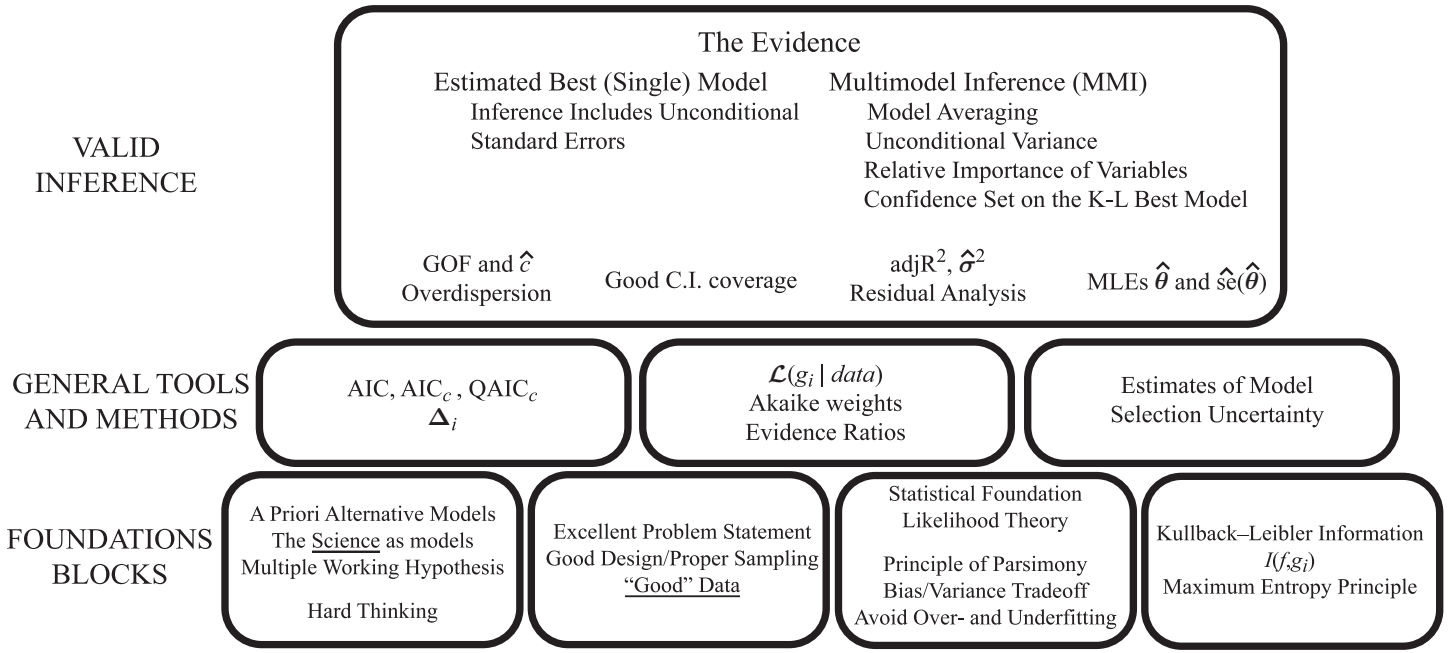

FIGURE 8.1. Schematic diagram of the information-theoretic approach. The evidence for the alternative hypotheses, each represented by mathematical models, and the analysis results are provided by the methods and quantities indicated in the top box. This information results from the use of the general methods in three linked, general tool boxes, which rest on the concepts and deep theory in four basic foundation blocks.

8.12 Final Thoughts

At a conceptual level, reasonable data and a *good* model allow a separation of "information" from "noise." Here, information relates to the structure of relationships, estimates of model parameters, and components of variance. Noise then refers to the residuals; variation left unexplained. We can use the information extracted from the data to make proper inferences.

Summary

We want an approximating model that minimizes information loss $I(f, g)$ and properly separates noise (noninformation, or entropy) from structural information. The philosophy for this separation is the principle of parsimony; the conceptual target for such partitioning is Kullback–Leibler information; and the tactic for selection of a best model is an information criterion (e.g., AIC, AIC_c, QAIC_c, or TIC). The notion of data-based model selection and resulting inference is a very difficult subject, but we do know that substantial uncertainty about the selected model can often be expected and should be incorporated into estimates of precision.

Still, model selection (in the sense of parsimony) is the critical issue in data analysis. In using the more advanced methods presented here, model selection can be thought of as a way to compute Akaike weights. Then one uses one or more models in the set as a way to make robust inferences from the data (MMI). More research is needed on the quantification of model uncertainty, measures of the plausibility of alternative models, ways to reduce model selection bias, and ways to provide effective measures of precision (without being conditional on a given model). Confidence intervals with good achieved levels should be a goal of inference following data-based model selection.

Information-theoretic methods are relatively simple to understand and practical to employ across a very wide class of empirical situations and scientific disciplines. The information-theoretic approach unifies parameter estimation and model selection under an optimization framework, based on Kullback–Leibler information and likelihood theory. With the exception of the bootstrap, the methods are easy to compute by hand if necessary (assuming that one has the MLEs, maximized log-likelihood values, and $\widehat{\text{var}}(\hat{\theta}_i \mid g_i)$ for each of the R models). Researchers can easily understand the information-theoretic methods presented here; we believe that it is *very* important that researchers understand the methods they employ.

http://www.springer.com/978-0-387-95364-9

Model Selection and Multimodel Inference
A Practical Information-Theoretic Approach
Burnham, K.P.; Anderson, D.R.
2002, XXVI, 488 p., Hardcover
ISBN: 978-0-387-95364-9