Model Selection and Multimodel Inference

Second Edition

Springer
New York
Berlin
Heidelberg
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Kenneth P. Burnham   David R. Anderson

Model Selection and Multimodel Inference

A Practical Information-Theoretic Approach

Second Edition

With 31 Illustrations

Springer

Kenneth P. Burnham
David R. Anderson
Colorado Cooperative Fish
and Wildlife Research Unit
Colorado State University
Fort Collins, CO 80523-1484
USA

Cover Illustration: The cover was assembled from photos of the yellow-bellied toad (*Bombina variegata*) taken by Jonas Barandum as part of his Ph.D. program at the University of Zurich. These toads have individually identifiable patterns on their abdomen from a few weeks following metamorphosis that remain unchanged until death. Two pairs are duplicates—but which two?
Cover photographs by Dr. Jonas Barandum, St. Gallen, Switzerland. Cover design by Kenton Allred.

Library of Congress Cataloging-in-Publication Data
Burnham, Kenneth P.
Model selection and multimodel inference : a practical information-theoretic approach / Kenneth P. Burnham, David R. Anderson.—2nd ed.
p. cm.
Rev. ed. of: Model selection and inference. © 1998.
Includes bibliographical references (p.).
ISBN 0-387-95364-7 (alk. paper)
1. Biology—Mathematical models. 2. Mathematical statistics. I. Burnham, Kenneth P. Model selection and inference. II. Title.
QH323.5 B87 2002
570′.1′51—dc21 2001057677

ISBN 0-387-95364-7 Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 SPIN 10853081

www.springer-ny.com

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science+Business Media GmbH

To my mother and father, Lucille R. (deceased) and J. Calvin Burnham (deceased), and my son and daughter, Shawn P. and Sally A. Burnham

To my parents, Charles R. (deceased) and Leta M. Anderson; my wife, Dalene F. Anderson; and my daughters, Tamara E. and Adrienne M. Anderson

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 accumulation 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

About the Authors

Drs. Kenneth P. Burnham and David R. Anderson have worked closely together for the past 28 years and have jointly published 9 books and research monographs and 66 journal papers on a variety of scientific issues. Currently, they are both in the Colorado Cooperative Fish and Wildlife Research Unit at Colorado State University, where they conduct research, teach graduate courses, and mentor graduate students.

Ken Burnham has a B.S. in biology and M.S. and Ph.D. degrees in statistics. For 29 years post-Ph.D. he has worked as a statistician, applying and developing statistical theory in several areas of life sciences, especially ecology and wildlife, most often in collaboration with subject-area specialists. Ken has worked (and lived) in Oregon, Alaska, Maryland (Patuxent Wildlife Research Center), North Carolina (U.S. Department of Agriculture at North Carolina State University, Statistics Department), and Colorado (currently USGS-BRD). He is the recipient of numerous professional awards including Distinguished Achievement Medal from the American Statistical Association, and Distinguished Statistical Ecologist Award from INTECOL (International Congress of Ecology). Ken is a Fellow of the American Statistical Association.

David Anderson received B.S. and M.S. degrees in wildlife biology and a Ph.D. in theoretical ecology. He is currently a Senior Scientist with the Biological Resources Division within the U.S. Geological Survey and a professor in the Department of Fishery and Wildlife Biology. He spent 9 years at the Patuxent Wildlife Research Center in Maryland and 9 years as leader of the Utah Cooperative Wildlife Research Unit and professor in the Wildlife Science Department at Utah State University. He has been at Colorado State University since 1984. He is the recipient of numerous professional awards for scientific and academic contributions, including the Meritorious Service Award from the U.S. Department of the Interior.

Glossary

Notation and abbreviations generally used are given below. Special notation for specific examples can be found in those sections.

AIC	Akaike's information criterion.
AIC_{min}	The estimate of relative, expected K-L information for the best model in the set, given the data. For example, given the models $g_1, g_2, \ldots, g_R$ and the data x, if the information criterion is minimized for model g_6, then $min = 6$, signifying that AIC_6 is the minimum over $\text{AIC}_1, \ldots, \text{AIC}_R$. The minimum AIC is a random variable over samples. This notation, indicating the index number in $\{1, 2, \ldots, R\}$ that minimizes expected K-L information, also applies to AIC_c, QAIC_c, and TIC.
AIC_{best}	In any set of models, one will be the best expected K-L model, hence the actual best AIC model. The model for which $E_f(\text{AIC})$ is minimized is denoted by the index *best*, whereas *min* is a random variable (like $\hat{\theta}$), *best* is fixed (like θ). This value can be determined using Monte Carlo methods. This "best" model is the same model over all possible samples (of which we have only a single sample). This notation also applies to AIC_c, QAIC_c, and TIC.
AIC_c	A second-order AIC, necessary for small samples.
Akaike weights	The relative likelihood of the model, given the data. These are normalized to sum to 1, are denoted by w_i, and interpreted as probabilities.

best	An index to denote the theoretically best fitted model; this model is best in the sense of expected K-L information, given the data. Such a best model can be found from Monte Carlo methods and represents a statistical expectation. For example, consider the set E(AIC_i), where $i = 1, 2., \ldots, R$. Then, the model where E(AIC_i) is minimized is denoted by AIC_{best}. AIC, AIC_c, QAIC_c, or TIC could be used in this context.
Bias	(of an estimator) Bias = $\text{E}(\hat{\theta}) - \theta$.
BIC	Bayesian information criterion (Akaike 1978a,b; Schwarz 1978), also termed SIC in some literature.
c	A simple variance inflation factor used in quasi-likelihood methods where there is overdispersion of count data (e.g., extra binomial variation).
Δ_i	AIC differences, relative to the smallest AIC value in the set of R models. Hence, AIC values are rescaled by a simple additive constant such that the model with the minimum AIC value has $\Delta_i = 0$. Formally, $\Delta_i = \text{AIC}_i - \text{AIC}_{min}$. These values are estimates of the expected K-L information (or distance) between the selected (best) model and the ith model. These differences apply to AIC, AIC_c, QAIC_c, or TIC.
Δ_p	A "pivotal" value, analogous to $(\theta - \hat{\theta})/\hat{s}\text{e}(\hat{\theta})$; $\Delta_p = \text{AIC}_{best} - \text{AIC}_{min}$.
df	Degrees of freedom as associated with hypothesis testing. The df is the difference between the number of parameters in the null and alternative hypotheses in standard likelihood ratio tests.
$\text{E}(\hat{\theta})$	The statistical expectation of the estimator $\hat{\theta}$.
Estimate	The computed value of an estimator, given a particular set of sample data (e.g., $\hat{\theta} = 9.8$).
Estimator	A function of the sample data that is used to estimate some parameter. An estimator is a random variable and is denoted by a "hat" (e.g., $\hat{\theta}$).
Evidence ratio	The relative likelihood of model i versus model j (e.g., $\mathcal{L}(g_i\|data)/\mathcal{L}(g_j\|data)$, which is identical to w_i/w_j).
$f(x)$	Used to denote "truth" or "full reality," the process that produces multivariate data x. This conceptual probability distribution is often considered to be a mapping from an infinite-dimensional space.
$g_i(x)$	Used to denote the set of candidate models that are hypothesized to provide an adequate approximation for the distribution of empirical data. The expression $g_i(x \mid \theta)$ is used when it is necessary to clarify that the function involves parameters θ.

	Often, the parameters have been estimated; thus the estimated approximating model is denoted by $g_i(x \mid \hat{\theta})$. Often, the set of R candidate models is represented as simply $g_1, g_2, \ldots, g_R$. Also, $\hat{g}_i = g_i(x\|\hat{\theta})$.
Global model	A highly parameterized model containing the variables and associated parameters thought to be important as judged from an a priori consideration of the problem at hand. When there is a global model, all other models in the set are special cases of this global model.
K	The number of estimable parameters in an approximating model.
K-L	Kullback–Leibler distance (or discrepancy, information, number).
LRT	Likelihood ratio test.
LS	Least squares method of estimation.
$\mathcal{L}(\theta \mid x, g)$	Likelihood function of the model parameters, given the data x and the model g.
$\mathcal{L}(g_i\|x)$	The discrete likelihood of model g_i, given the data x.
$\log(\cdot)$	The natural logarithm ($\log_e$).
$\text{logit}(\theta)$	The logit transform: $\text{logit}(\theta) = \log(\theta/(1-\theta))$, where $0 < \theta < 1$.
g_i	Shorthand notation for the candidate models considered.
min	An index to denote the fitted model that minimizes the information criterion, given the data. Then, model g_{min} is the model selected, based on minimizing the appropriate criterion, given the data. AIC, AIC_c, QAIC_c, or TIC could be used in this context.
ML	Maximum likelihood method of estimation.
MLE	Maximum likelihood estimate (or estimator).
n	Sample size. In some applications there may be more than one relevant sample size (e.g., in random effects models).
Parsimony	The concept that a model should be as simple as possible concerning the included variables, model structure, and number of parameters. Parsimony is a desired characteristic of a model used for inference, and it is usually defined by a suitable tradeoff between squared bias and variance of parameter estimators. Parsimony lies between the evils of under- and over-fitting.
Precision	A property of an estimator related to the amount of variation among estimates from repeated samples.

$\propto$	A symbol meaning "proportional to."
QAIC or QAIC_c	Versions of AIC or AIC_c for overdispersed count data where quasi-likelihood adjustments are required, hence $\hat{c}$ used.
π_i	Model selection probabilities (or relative frequencies), often from Monte Carlo studies or the bootstrap.
R	The number of candidate models in the set; $i = 1, 2, \ldots, R$. One of these models is the estimated best model (i.e., in the sense of a specific model $g(x\|\hat{\theta})$, where the model parameters have been estimated) for the data at hand (g_{min}). One model (possibly the same model) is the theoretically best model (g_{best}) to use as a basis for inference from the data.
τ_i	Prior probability of model i. Also used to cope with model redundancy (Section 4.6).
θ	Used to denote a generic parameter vector (such as a set of conditional survival probabilities S_i).
$\hat{\theta}$	An estimator of the generic parameter θ.
θ_0	The optimal parameter value in a given model g, given a fixed sample size, but ignoring estimation issues (see Section 7.1). This is the value that minimizes K-L information, given the model structure.
TIC	Takeuchi's information criterion.
w_i	Akaike weights. Used with any of the information criteria that are estimates of expected Kullback–Leibler information (AIC, AIC_c, QAIC, TIC). The w_i sum to 1 and may be interpreted as the probability that model i is the actual expected K-L best model for the sampling situation considered.
$w_+(j)$	Sum of Akaike weights over all models that include the explanatory variable j. These sums are useful in variable-selection problems where one wants a measure of relative importance of the explanatory variables and in computing estimates that are robust to model selection bias.
χ^2	A test statistic distributed as chi-squared with specified degrees of freedom df. Used here primarily in relation to a goodness-of-fit test of the global model in analyzing count data.
$\approx$	Approximately equal to.
$\sim$	Distributed as.