UNDERSTANDING INFORMATION CRITERIA FOR SELECTION AMONG CAPTURE-RECAPTURE OR RING RECOVERY MODELS

David R. Anderson and Kenneth P. Burnham
Colorado Cooperative Fish and Wildlife Research Unit
Room 201 Wagar Building, Colorado State University
Fort Collins, Colorado USA 80523-1484

Short Title: Understanding Information Criteria

Summary

We provide background information to allow a heuristic understanding of two types of criteria used in selecting a model for making inferences from ringing data. The first type of criteria (e.g., AIC, AIC_c, $QAIC_c$ and TIC) are estimates of (relative) Kullback-Leibler information or distance and attempt to select a good approximating model for inference, based on the Principle of Parsimony. The second type of criteria (e.g., BIC, MDL, HQ) are "dimension consistent" in that they attempt to consistently estimate the dimension of the true model. These latter criteria assume that a true model exists, that it is in the set of candidate models and that the goal of model selection is to find the true model, which in turn requires and that sample size is very large. The Kullback-Leibler based criteria do not assume a true model exists, let alone that it is in the set of models being considered. Based on a review of these criteria we recommend use of criteria that are based on Kullback-Leibler information in the biological sciences.

1. Introduction

The analysis of ringing data requires models to link the unknown parameters to the data and the assumed structure of the model. Models of ringing data include parameters of primary interest (e.g., survival probabilities, ϕ and S) as well as "nuisance" parameters which are of secondly interest (e.g., the sampling probabilities, p, f and r). Each of these types of parameters can be constant, year- or age- or sex- or area-dependent; thus a large number of possible models can be formulated for any particular data set (see Lebreton et al.[1]). But, "*which model should be used?"*

Given a model, there exists a great deal of theory for making estimates of model parameters, based on the empirical data. Estimators of precision (standard errors, coefficients of variation and confidence intervals) can be derived, given a model. Likelihood and least squares theory provide rigorous, omnibus inference methods if the model is given. Thus, a central problem in the analysis of ringing data is which model to use for making inferences from the data – this is the *model selection* problem. One then uses the data to help select an appropriate model as well as make estimates of the unknown parameters.

Under the frequentist paradigm for model selection, there are three general approaches: (I) optimization of some selection criterion, (II) tests of hypotheses, and (III) ad hoc methods. One has a further classification within the selection criteria optimization approach: (1) criteria based on some form of mean squared error (MSE) or mean squared prediction error (MSPE) (examples here include C_p (Mallows[2]) and PRESS (Allen[3])), (2) criteria that are estimates of relative Kullback-Leibler (K-L) information and (3) criteria that are consistent estimators of K, the dimension of the "true model." The objective of this paper is to explain, compare and contrast the fundamental basis for these latter two classes of selection criteria.

We use $f(x)$ to denote the concept of "full truth" or "reality" and $g_i(x)$ ($i = 1, \ldots, M$) to denote a set of approximating models, where x represents the (multivariate) data. In practice, any approximating model $g(x)$ has K parameters (θ) that must be estimated from the finite data set x. In applied problems in the analysis of ringing data, K might range from as few as 2-4 to 40-80 parameters, or perhaps more, within the set of M approximating models.

The K-L information or "distance" between f and an approximating model g is defined as

$$I(f, g) = \int f(x) \log_e \left(\frac{f(x)}{g(x)} \right) dx \qquad \text{or} \qquad I(f, g) = \sum_{i=1}^{k} p_i \log_e \left(\frac{p_i}{\pi_i} \right)$$

for continuous and discrete distributions, respectively. For discrete distributions there are k possible outcomes of the underlying random variable; the true probability of the i^{th} outcome is given by p_i, while the $\pi_1, \ldots, \pi_k$ constitute the probabilities from the approximating model (probability distribution). Kullback and Leibler[4]) developed $I(f, g)$ from "information theory" (see Guiasu[5], Cover and Thomas[6]) as they sought to define rigorously what R. A. Fisher meant by "information" in his concept of sufficient statistics. The quantity $I(f, g)$ measures the "information" lost when g is used to approximate f (truth). If $f \equiv g$, then $I(f, g) = 0$, as there is no information lost when the model reflects truth perfectly. In real applications, some information will invariably be lost when a model is used to approximate full reality about ringed populations, thus $I(f, g) > 0$. $I(f, g)$ can also be thought of as a "distance" between f and g. We will use both meanings as they both offer worthwhile insights.

Although derived along very different lines, K-L information is the negative of Boltzmann's *entropy*, a crowning achievement of 19th century science (see Broda[7]). Moreover, a deep, fundamental result of information theory (developed in the mid 20th century) is that "information" is related to the logarithm of a probability (discrete case) or of a probability distribution function (see e.g., Cover and Thomas[6]). Given this result, then the cross information between f and g, in the sense of g approximating f, is best measured, by far, by $I(f, g)$ (see also Kapur and Kesavan[8]): there is no justified information theory-based competing measure to $I(f, g)$.

We seek a model that loses as little information as possible or one that is the shortest distance away from truth. We also assume (correctly, we believe) that no model in the set of models considered is true; hence, selection of a best approximating model must be our goal. As a consequence of the above, this conceptual goal must be taken as equivalent to minimizing $I(f, g)$. Operationally, K-L cannot be used directly in model selection as it requires knowledge of both f (truth) and the parameters in $g(x)$ ($\equiv g(x \mid \theta)$). In the material below we will

motivate the concept that relative K-L information can be estimated from the data based on the maximized log-likelihood function.

K-L information can be expressed as a difference between two statistical expectations (thinking of x as a random variable) with respect to the distribution f,

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

The first expectation, $\mathrm{E}_f\,[\log(f(x))]$, is an unknown constant (it will not vary by approximating model) that depends only on the unknown truth, f. Therefore, treating this unknown term as a constant, only a measure of *relative* information is possible (Bozdogan[9], Kapur and Kesavan[8] p. 155). Clearly, if one could estimate the second expectation, $\mathrm{E}_f\,[\log(g(x \mid \theta))]$, one could estimate $I(f, g)$, up to an additive constant (namely $\mathrm{E}_f\,[\log(f(x))]$) that does not depend on the assumed model,

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

note that the expectations needed here are taken with respect to unknown truth, f. These expressions would provide the basis for estimated *relative* distance between the two models f and g, if only one could estimate $\mathrm{E}_f\,[\log(g(x \mid \theta))]$ – this is Akaike's achievement.

In the following material we assume unknown model parameters are estimated using Fisher's maximum likelihood method and we assume there is a log-likelihood function $\log(\mathcal{L}(\theta \mid x)) = \log(g(x \mid \theta))$ associated with each probability model in the set of M approximating models.

2. Information Criteria That Are Estimators of K-L Information

Akaike's[10] seminal paper proposed the use of the Kullback-Leibler information as a fundamental basis for model selection. Akaike[10–14] showed that the critical term is the relative K-L information and this quantity can be *estimated* from the empirical data and the maximized log-likelihood function. Akaike's finding of a relation between the relative K-L information and the maximized log-likelihood has allowed major practical and theoretical advances in model selection and the analysis of complex data sets (see Stone[15], Shibata[16], Bozdogan[9], and deLeeuw[17]). For the *iid* case, there is a relationship with cross validation principles (Stone[18,19] and Stoica et al.[20]).

2.1 Heuristic Background

The statistical theory underlying the derivations is quite deep (Burnham and Anderson[21] give it in detail), here we will provide only an overview. One key point is that we have to estimate relative K-L distance based on $\hat{g} \equiv g(x \mid \hat{\theta})$, but unmodified $I(f, \hat{g})$ turns out to not be a suitable selection criterion because it has two strong sources of bias. First, the use of the MLE $\hat{\theta}$ is itself a source of downward bias; second, there is an additional downward bias that depends strongly on K. Hence, we determine $\mathrm{E}_f\,[I(f, \hat{g})]$ relative to our target K-L selection criterion and do a bias correction on what amounts to $I(f, \hat{g})$ minus the Constant noted above. Heuristically, this gives us a bias-corrected, estimated relative K-L distance.

Although AIC was published in 1973, it is convenient to introduce the concept using Takeuchi's[22] result for the general asymptotic relationship between the target criterion, $I(f, g)] - \text{Constant}$, and $\mathrm{E}_f\,[\log(g(x \mid \hat{\theta}))]$ which is the expected empirical maximized log-likelihood. Here, $\hat{\theta}$ is the MLE of the K model parameters under model g, based on data x. Takeuchi[22] obtained

$$-\mathrm{E}_f\,[\log(\mathcal{L}(\hat{\theta} \mid x))] = I(f, g) - \text{Constant} - \text{trace}\left(J(\theta_o)[I(\theta_o)]^{-1}\right).$$

This is equivalent to

$$-\mathrm{E}_f\,[\log(\mathcal{L}(\hat{\theta} \mid x))] + \mathrm{trace}\left(J(\theta_o)[I(\theta_o)]^{-1}\right) = I(f, g) - \text{Constant}.$$

An unbiased estimator of $\mathrm{E}_f\,[\log(\mathcal{L}(\hat{\theta} \mid x))]$ is simply $\log(\mathcal{L}(\hat{\theta} \mid x))$ itself. Hence, the issue of a computable K-L based criterion reduces to computing (i.e., estimating) the above matrix trace term. Here, J and I are K by K matrices based, respectively, on first and second partial derivatives of $\mathrm{E}_f\,[\log(g(x \mid \theta))]$ with respect to θ. These matrices are related to, but not identical to, the Fisher information matrix. For count data (which applies to ringing data), θ_o is computed as the MLE when the data, x, are (suitably) replaced by their expectations $\mathrm{E}_f(x)$ taken with respect to true $f(x)$. This value of $\theta = \theta_o$ is the value that minimizes $I(f, g)$ for the model g over all possible values of θ.

Derivation of the above asymptotically justified formula does not require any assumption that one or more of the approximating models is "good" or even close to truth. This formula is a general result and can be used as a criterion for model selection (called TIC for Takeuchi's Information Criterion). However, if $f(x) = g(x)$, then $J(\theta_o) = I(\theta_o)$ (and then they are the Fisher information matrix) hence $J(\theta_o)[I(\theta_o)]^{-1}$ is the K by K identity matrix, and the trace of that matrix is just K (the dimension of the approximating model). It also holds that if g is a good approximation to f, then $\mathrm{trace}\left(J(\theta_o)[I(\theta_o)]^{-1}\right)$ is well approximated by K. Thus, if at least one of the approximating models is relatively good then the matrix trace of $J(\theta_o)[I(\theta_o)]^{-1}$ can be safely estimated simply by K over the set of models used (there is more to this matter than we can give here, see also Burnham and Anderson[21]). This leads directly to Akaike's Information Criterion (AIC). The alternative is to estimate the elements of the matrices $J(\theta_o)$ and $I(\theta_o)$ from the data, however the sampling variance of the resultant estimated matrix trace may often be substantial.

Thus, in the sense of parsimony, letting $\mathrm{trace}\,J(\theta_o)[I(\theta_o)]^{-1} = K$ is a justified, and almost necessary, alternative (unless the sample size is very large). Then a suitable estimator of the relative K-L information (i.e., $\mathrm{E}_f\,[\log(g(x \mid \theta_o))]$, or more precisely an estimator of, $I(f, g) - \text{Constant}$) is

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

Akaike[10] then defined "*an information criterion*" (AIC) by multiplying through by -2 ("taking historical reasons into account"),

$$\mathrm{AIC} = -2\log(\mathcal{L}(\hat{\theta} \mid x)) + 2K\,.$$

This has became known as "*Akaike's Information Criterion*" or AIC. AIC is an approximation whereby the matrix trace of the product of the matrices J and I^{-1} is replaced by a simple, known scalar, K.

Here it is important to note that AIC has a strong theoretical underpinning, based on information theory, K-L information and ML theory. AIC is useful for both nested and non-nested models. Akaike's inferential breakthrough was finding that a predictive expectation of the log-likelihood could be used to estimate the relative K-L information. The constant term ("Constant," above) involving $f(x)$ is independent of the data and models and therefore drops out of the model selection criterion leaving AIC defined without specific reference to a "true model" (Akaike[13] p. 13). Thus, one should select the model that yields the smallest value of AIC, because this model is estimated to be "closest" to the unknown truth, among the candidate models considered.

The K-L information can be made smaller by adding more structure, via parameters, to the approximating model g. However, when the added parameters are estimated (rather than being known or "given"), further uncertainly is added to the *estimation* of the relative K-L information. Thus, for a fixed sample size, the addition of estimated parameters to a poor fitting model $\hat{g}_i$, thus getting a different model, $\hat{g}_j$, will allow the expanded fitted model to be closer to f. However, at some point, the addition of still more estimated parameters (to what may already be a suitable fitted model) will have the opposite effect and the *estimate* of the relative K-L information will increase. This fact represents the trade-off between bias and variance, or the trade-off between under-fitting and over-fitting the sample data, that is fundamental to the Principle of Parsimony.

The AIC criterion may perform poorly if there are too many parameters in relation to the size of the sample (Sugiura[23], Sakamoto et al.[24]). Based on the initial work of Sugiura[23], Hurvich and Tsai[25] derived a small-sample (second order) expression which led to a refined criterion denoted as AIC_c ,

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

or

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1} ,$$

where n is sample size (see Sugiura[23]). In ringing studies it is not always clear as to what constitutes "sample size." Conservatively, n is the number of distinct birds ringed during a study. If n is large with respect to K, then the second order correction is negligible and AIC should perform well.

While AIC_c was derived under Gaussian assumptions, Burnham et al. [26]) found this second order expression to perform well in product multinomial models for open population capture-recapture. Generally, we advocate the use of AIC_c when the ratio n/K is small (say < 40). AIC_c is a better (second order) approximation to the relative K-L information and we recommend its general use when sample size is not large relative to the number of parameters being estimated. If the ratio n/K is sufficiently large, then AIC and AIC_c are similar and will tend to select the same approximating model.

In capture-recapture and ring recovery data analysis there is often overdispersion (see Lebreton et al.[1]) and then modified criteria are appropriate,

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

and

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

where a variance inflation factor (c) is estimated from the goodness-of-fit chi-square statistic (χ^2) and its degrees of freedom (df), $\hat{c} = \chi^2/df$. The estimate of c should come from the highest dimensioned model in the set of candidates (the global model). Of course, when no overdispersion exists, c = 1; then the formulae for QAIC and $QAIC_c$ reduce to AIC and AIC_c, respectively (see Anderson et al.[27]).

2.2 Summary

All of AIC, AIC_c, QAIC, $QAIC_c$ and TIC are estimates of the relative K-L information between f and each of the M approximating models $g_i(x)$. These criteria were derived based on the concept that truth is very complex and that no "true model" exists (or, at least, it is very high dimensional). Thus, one could only *approximate* truth with a model, say $g(x)$. Given a good list of candidate models for the data, one could estimate which approximating model was best (closest to unknown truth), among those candidates considered, given the data and their sample size (Hurvich and Tsai[28]). Also, the size (K) of the models should depend on sample size; as more data become available more elaborate models are justified. The basis for these criteria thus seem reasonable in the biological sciences if some good approximating models are in the *a priori* set of candidates.

Akaike's contribution is more than merely computing AIC values as estimates of the K-L information for each model and selecting the model that minimizes this quantity. He also suggests the importance of *a priori* modeling to incorporate the science concerning what is known or hypothesized. Thus, a small set of candidate models are derived after serious thought about the question to be addressed, what is known from the literature, and what is hypothesized. These considerations are to be made carefully and before data analysis begins. This process has not been common practice in the past and needs an emphasis in the analysis of bird ringing data.

3. Criteria That Are Consistent for K, the Dimension of the True Model

3.1 History

Following Akaike's pioneering derivation of AIC, people noticed a heuristic interpretation that was both interesting and sometimes very misleading. The first term in AIC,

$$AIC = -2\log(\mathcal{L}(\hat{\theta} \mid x)) + 2K,$$

is a measure of lack of model fit, while the second term ($2K$) can be interpreted as a "penalty" for increasing the size of the model (the penalty enforces parsimony in the number of parameters). This heuristic explanation does not do justice to the much deeper theoretical basis for AIC (i.e., AIC is an estimator of relative K-L information, and K-L information is a compelling measure of the closeness of a model to truth). This heuristic interpretation led some statisticians to consider "alternative" penalty terms whereby they focused on consistent (asymptotically unbiased, variance tending to 0) estimation of K, the dimension of the (assumed) $true$ model $f(x)$. Such criteria are termed "dimension consistent" (Bozdogan[9]). Part of the required philosophy behind such criteria is that the true model is in the set of models considered and the goal of model selection is to select this true model with probability 1 as sample size gets large. Both of these features are contrary to the philosophy behind AIC (and unrealistic, we think).

The best known of the "dimension consistent" criteria was derived by Schwarz[29] in a Bayesian context and is termed BIC for Bayesian Information Criterion; it is simply,

$$\text{BIC} = -2\log(\mathcal{L}(\hat{\theta} \mid x)) + K \cdot \log(n).$$

BIC was derived in a fully Bayesian context with prior probability $1/M$ on each of M models and very vague priors on the parameters θ in each model in the set. BIC has been widely used in several applied fields.

Rissanen[30] proposed a criterion he called Minimum Description Length (MDL), based on coding theory, another branch of information theory (also see Rissanen[31]). While the derivation and its justification are difficult to follow without a strong background in coding theory, his criterion is equivalent to BIC. Hannan and Quinn[32] derived a criterion (HQ) for model selection whereby the penalty term was

$$c \cdot \log(\log(n)),$$

where c is a constant > 2 (see Bozdogan[9] p. 359). This criterion, while often cited, has seen little use in practice. Bozdogan[9] proposed a criterion he called CAICF (C denoting "consistent" and F denoting the use of the Fisher information matrix, $\mathcal{I}$),

$$\text{CAICF} = -2\log(\mathcal{L}(\hat{\theta} \mid x)) + K\left(\log(n) + 2\right) + \log \mid \mathcal{I}(\hat{\theta}) \mid ,$$

where $\log \mid \mathcal{I}(\hat{\theta}) \mid$ is the natural logarithm of the determinant of the estimated Fisher information matrix. For large n, CAICF behaves like BIC; however, CAICF is not invariant to transformations of the parameters and this would seem to be a substantial defect. Bozdogan[33] has introduced a criterion denoted as Informational Complexity Criterion (ICOMP), defined as

$$-2\log(\mathcal{L}(\hat{\theta} \mid x)) + 2C(\hat{\Sigma}_{model}),$$

where C is a complexity measure and Σ_{model} is the variance-covariance matrix of the parameters estimated under the model. Several approaches and extensions are considered, however, these details would take us too far afield here.

3.2 Summary

The dimension consistent criteria are based on the assumption that an exactly "true model" exists *and* that it is one of the candidate models being considered. Implicit are the assumptions that truth is of fairly low dimension (i.e., $K =$ 1-5 or so) and that K is fixed as sample size increases. Finally, these criteria assume effects are either large or 0; tapering effect sizes complicate the matter of K (Speed and Yu[34]). Here, the criteria are derived to provide a consistent estimator of the order or dimension (K) of this "true model" and the probability of selecting this "true model" approaches 1 as sample size increases. Bozdogan[9] provides a nice review of many of these dimension consistent criterion and Shibata[35] gives a more recent and technical treatment.

4. Summary and Discussion

We question the concept of a low dimensional “true model" in the biological sciences. Even if a “true model" existed, surely it would be of high dimension (e.g., due to individual heterogeneity, in addition to a host of other sources of variability and various interactions). The dimension consistent criteria are based on the asymptotics that as $n \rightarrow \infty$, K remains fixed and small. In reality in the biological sciences, as n increases substantially, K must also increase. Relatively few people seem to be aware of the fundamental differences in the basis and assumptions for the dimension consistent criteria. As Reschenhofer[36] notes, regarding AIC vs. BIC, they “… are often employed in the same situations which is in contrast to the fact that they have been designed to answer different questions" (also see Potscher[37], Hurvich and Tsai[28,38], Shibata[35]). In the biological and social sciences and medicine, we argue that the AIC-type criteria are reasonable for the analysis of empirical data. In contrast, we cannot recommend the use of the dimension consistent criteria in model selection in the biological sciences; these criteria seem to be based on several assumptions that are not valid and hence have unrealistic goals.

Not withstanding our objections above, the sample sizes required to achieve the benefits of consistent estimation of model order (K) are often very large by any usual standard. In Monte Carlo examples we have studied, we have seen the need for sample sizes in the thousands or much more before the consistent criteria begin to point reliably to the true K, or “true model," with a high probability (even when the generating model is of low dimension). In cases where n was very large, say 100,000 or a million, one might merely examine the ratios $\hat{\theta}/\widehat{\text{se}}(\hat{\theta})$ to decide on the parameterization, with little regard for the principle of parsimony (because it is the true model being sought and, under the dimension consistent concept, it is assumed to be in the set of candidate models).

It should be emphasized that the dimension consistent criteria are not linked directly to K-L information and are “information theoretic" only in the weakest sense (perhaps a misnomer). Except for historical precedent, they should not be termed “… information criteria." Instead, their motivation veered toward consistent estimation of the order (K) of the supposed “true model" (often in an autoregression time series context) by employing alternative penalty terms to achieve “consistency."

When sample size is less than very large, these dimension consistent criteria tend to select under-fitted models with the attendant substantial bias, overestimated precision and associated problems in inference (Burnham et al.[39,40], Anderson et al.[41]). Shibata[42] argues convincingly that over-fitting is less problematic than under-fitting and that AIC attempts a “balance" between these undesirable alternatives (see Burnham et al.[39]).

The dimension consistent criteria might find use in the physical sciences where a true model might well exist, is contained in the list of candidate models, is of low dimension, and where sample size is quite large (perhaps thousands or tens of thousands, or more). Even in (artificial) cases where a true model exists and it is contained in the list of candidates, AIC might frequently have better inferential properties than the dimension consistent criteria unless the sample size is very large.

People have often used Monte Carlo methods to study the various criteria and this has been the source of confusion in many cases (Rosenblum[43]). In Monte Carlo studies, one *knows* the generating model (both its exact form and its parameters), it is usually somewhat simple with several “big" effects and few if any tapering effects, and the data generating model is nearly always included in the list of candidate models (the $g_i(x)$). Also, attention is often focused on which criterion (AIC vs. BIC) most often selects this true model. These conditions are those for which BIC is intended and AIC is not intended. It is thus not surprising that in this artificial situation the dimension consistent criteria may perform well relative to AIC, especially if the order of the true model is quite low, there are no tapering effects, many models that are too general are included, and the sample size is large. This situation is, however, quite unlike that faced by biologists in analyzing ringing data.

More informative and realistic Monte Carlo studies would employ a range of tapering effects and a high dimensional generating model that is *not* in the list of candidate models. Then, attention can be focused on the utility of the selected best approximating model and the validity of inferences drawn from it (as full truth is known and can serve as a basis for comparison). Within this Monte Carlo framework we have found that the dimension consistent criteria perform poorly in open population capture-recapture models even in the case where K is small, but the parameters reflect a range of effect sizes (Anderson et al. in press). In contrast, AIC performed well.

5. References

1. Lebreton, J-D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monogrographs* **62,** 67-118.

2. Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **12**, 591–612.

3. Allen, D. M. (1970). Mean square error of prediction as a criterion for selecting variables. *Technometrics* **13**, 469-475.

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

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

6. Cover, T. M., and Thomas, J. A. (1991). *Elements of information theory*. John Wiley and Sons, New York. 542 pp.

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

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

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

10. Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. Pages 267-281 in *Second international symposium on information theory*. B. N. Petrov and F. Csaki, (editors). Akademiai Kiado, Budapest.

11. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control AC* **19**, 716-723.

12. Akaike, H. (1977). On entropy maximization principle. Pages 27-41 in P. R. Krishnaiah (editor). *Applications of statistics*. North-Holland, Amsterdam.

13. Akaike, H. (1985). Prediction and entropy. in *A celebration of statistics*. Pages 1-24. inA. C. Atkinson and S. E. Fienberg (editors), Springer, New York.

14. Akaike, H. (1994). Implications of the informational point of view on the development of statistical science. Pages 27-38 in H. Bozdogan, (editor). Engineering and Scientific Applications,
Vol. 3, Proceedings of the First US/Japan Conference on the *Frontiers of statistical modeling: an informational approach*. Kluwer Academic Publishers, Dordrecht, Netherlands.

15. Stone, M. (1982). Local asymptotic admissibility of a generalization of Akaike's model selection rule. *Annals of the Institute for Statistical Mathematics,* (Part a) **34**, 123-133.

16. Shibata, R. (1983). A theoretical view of the use of AIC. Pages 237-244 in O. D. Anderson (editor). *Time series analysis: theory and practice*. Elsevier Scientific Publ., North-Holland.

17. deLeeuw, J. (1992). Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. Pages 599-609 in S. Kotz and N. L. Johnson (editors). *Breakthroughs in statistics,* Vol. 1. Springer-Verlag, London.

18. Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society,* Series B **39**, 111-147.

19. Stone, M. (1977). An asymptotic equivalence of choice of model by cross validation and Akaike's criterion. *Journal of the Royal Statistical Society,* Series B **39,** 44-47.

20. Stoica, P, Eykhoff, P., Janssen, P., and Söderström, T. (1986). Model-structure selection by cross-validation. *International Journal of Control* **43**, 1841-1878.

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

22. Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku* (Mathematic Sciences) **153**, 12-18. (In Japanese).

23. Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods* **A7**, 13-26.

24. Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). *Akaike information criterion statistics.* KTK Scientific Publishers, Tokyo.

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

26. Burnham, K. P., Anderson, D. R., and White, G. C. (1994). Evaluation of the Kullback–Leibler discrepancy for model selection in open population capture-recapture models. *Biometrical Journal* **36**, 299-315.

27. Anderson, D. R., Burnham, K. P., and White, G. C. (1994). AIC model selection in overdispersed capture-recapture data. *Ecology* **75**, 1780-1793.

28. Hurvich, C. M., and Tsai, C-L. (1994). Autoregressive model selection in small samples using a bias-corrected version of AIC. Pages 137-157 in H. Bozdogan, (editor). Engineering and Scientific Applications, Vol. 3, Proceedings of the First US/Japan Conference on the *Frontiers of statistical modeling: an informational approach.* Kluwer Academic Publishers, Dordrecht, Netherlands.

29. Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464.

30. Rissanen, J. (1989). *Stochastic complexity in statistical inquiry.* World Scientific, Singapore.

31. Rissanan, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory* **42**, 40-47.

32. Hannan, E. J., and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B* **41**, 190-195.

33. Bozdogan, H. (199_). Informational complexity criteria for regression models. *Computational Statistics and Data Analysis* **__, __-__.** (submitted)

34. Speed, T. P., and Yu, B. (1993). Model selection and prediction: normal regression. *Annals of the Institute of Statistical Mathematics* **1**, 35-54.

35. Shibata, R. (1998). *Statistical model selection*. Springer-Verlag, New York, NY.

36. Reschenhofer, E. (1996). Prediction with vague prior knowledge. *Communications in Statistics – Theory and Methods* **25**, 601-608.

37. Potscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory* **7**, 163-185.

38. Hurvich, C. M., and Tsai, C-L. (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics* **51,** 1077-1084.

39. Burnham, K. P., White, G. C., and Anderson, D. R. (1995a). Model selection in the analysis of capture-recapture data. *Biometrics* **51**, 888-898.

40. Burnham, K. P., Anderson, D. R., and White, G. C. (1995b). Selection among open populations capture-recapture models when capture probabilities are heterogeneous. *Journal of Applied Statistics* **22**, 611-624.

41. Anderson, D. R., Burnham, K. P., and White, G. C. (1998). Comparison of AIC and CAIC for model selection and statistical inference from capture-recapture studies. *Journal of Applied Statistics* 25:263-282.

42. Shibata, R. (1989). Statistical aspects of model selection. Pages 215-240 in J. C. Willems. (editor) *From data to model*. Springer-Verlag. London.

43. Rosenblum, E. P. (1994). A simulation study of information theoretic techniques and classical hypothesis tests in one factor ANOVA. Pages 319-346 in H. Bozdogan, (editor). Engineering and Scientific Applications, Vol. 3, Proceedings of the First US/Japan Conference on the *Frontiers of statistical modeling: an informational approach*. Kluwer Academic Publishers, Dordrecht, Netherlands.