SMR ms. of May 27, 2004

Kenneth P. Burnham
Room 201 Wagar Building
Colorado State University
Fort Collins, CO 80523
970 491-5396
970 491-1413 (FAX)
kenb@cnr.colostate.edu

David R. Anderson
Room 201 Wagar Building
Colorado State University
Fort Collins, CO 80523
970 491-5396
970 491-1413 (FAX)
anderson@cnr.colostate.edu

Multimodel Inference:
Understanding AIC and BIC in Model Selection

Kenneth P. Burnham
David R. Anderson
Colorado Cooperative Fish and Wildlife Research Unit (USGS-BRD)

The model selection literature has been generally poor at reflecting the deep foundations of AIC and at making appropriate comparisons to BIC. There is both a clear philosophy, a sound criterion based in information theory, and a rigorous statistical foundation for AIC. AIC can be justified as Bayesian using a “savvy" prior on models that is a function of sample size and the number of model parameters Furthermore, BIC can be derived as a non-Bayesian result. Therefore, arguments about using AIC versus BIC for model selection cannot be from a Bayes versus frequentist perspective. Deeper arguments must be explored and this is one of our objectives here. The philosophical context of what is assumed about reality, approximating models, and the intent of model-based

inference should determine whether AIC or BIC is used. Also, model selection must be more than just a search for, and then inference from, a single best model in a set: inference should reflect all models considered in the set. Various facets of such multimodel inference are presented here, particularly methods of model averaging.

1. INTRODUCTION

For a model selection context we assume there are data and a set of models and that statistical inference is to be model-based. Classically it is assumed that there is a single correct (or even true) or, at least, best model and that model suffices as the sole model for making inferences from the data. Whereas, the identity (and parameter values) of that model is unknown, it seems to be assumed that it can estimated, in fact well-estimated. Therefore, classical inference often involves a data-based search, over the model set, for (i.e., selection of) that single correct model (but with estimated parameters). Then inference is based on the fitted selected model as if it were the only model considered. Model selection uncertainty is ignored. This is considered justified because, after all, the single best model has been found. However, many selection methods used (e.g., classical stepwise selection) are not even based on an explicit criterion of what is a best model.

One might think the first step to improved inference under model selection would be to establish a selection criterion, such as AIC or BIC. However, we claim the first step is to establish a philosophy about models and data analysis and then find a suitable model selection criterion. The key issue of such a philosophy seems to center around one issue: are models ever true, in the sense of is full reality represented exactly by a model we can conceive and fit to the data; or are models merely approximations. Even minimally experienced practitioners of data analysis would surely say models are only approximations to full reality. Given this latter viewpoint, the issue is then really about whether the information (“truth") in the data, as extractable by the models in the set, is simple (a few big effects only) or complex (many

tapering effects). Moreover, there is a fundamental issue of seeking parsimony in model fitting: what “size" of fitted model can be justified given the size of the sample, especially in the case of complex data (we believe most real data are complex).

Model selection should be based on a well-justified criterion of what is the “best" model and that criterion should be based on a philosophy about models and model-based statistical inference, including the fact that the data are finite and “noisy." The criterion must be estimable from the data for each fitted model and the criterion must fit into a general statistical inference framework. Basically, this means model selection is justified and operates within either a likelihood or Bayesian framework, or within both frameworks. Moreover, this criterion must reduce to a number for each fitted model, given the data, and it must allow computation of model weights to quantify the uncertainty that each model is the target best model. Such a framework and methodology allows us to go beyond inference based on only the selected best model. Rather, we do inference based on the full set of models: multimodel inference. Very little of the extensive model selection literature goes beyond the concept of a single best model, often because it is assumed the model set contains the true model. This is true even for major or recent publications, e.g., Linhart and Zucchini (1986), McQuarrie and Tsai (1998), and Lahiri (2001).

Two well known approaches meet these conditions operationally: information-theoretic selection based on Kullback-Leibler (K-L) information loss and Bayesian model selection based on Bayes factors. Akaike's information criterion (AIC) represents the first approach. We will let the BIC approximation to the Bayes factor represent the second approach; exact Bayesian model selection (see e.g., Gelfand and Dey 1994) can be much more complex than BIC – too complex for our purposes here. The focus and message of our paper is on the depth of foundation underlying K-L information and AIC. Many people using, abusing or refusing AIC do not know its foundations, nor its current depth of development for coping with model selection uncertainty (multimodel inference). Moreover, understanding either AIC or BIC is enhanced by contrasting them; therefore, we will provide contrasts.

Another reason to include BIC here, despite AIC being our focus, is because using the BIC approximation to the Bayes factor we can show that AIC has a Bayesian derivation.

We will not give the mathematical derivations of AIC or BIC. Neither will we say much about the philosophy on deriving an a prior set of models. Mathematical and philosophical background for our purposes is given in Burnham and Anderson (2002). There is much other relevant literature that we could direct the reader to, for example, Akaike (1973, 1981) and deLeeuw (1992), about AIC, and Gelfand and Dey (1994), Gelman et al. (1995), Raftery (1995), Kass and Raftery (1995), Key et al. (1999), and Hoeting et al. (1999) about Bayesian model selection. For an extensive set of references we direct the reader to Burnham and Anderson (2002) and Lahiri (2001). We do not assume the reader has read all, or much, of this literature. However, we do assume the reader has a general familiarity with model selection including having encountered AIC and BIC, and arguments pro and con about which one to use (e.g., Weakliem 1999).

Our paper is organized around 5 sections. Section 2 is a careful review of K-L information, parsimony, AIC as an asymptotically unbiased estimator of relative, expected K-L information, AIC_c and TIC, scaling criterion values (Δ_i), the discrete likelihood of model *i*, given the data, Akaike weights, the concept of evidence, and measures of precision that incorporate model selection uncertainty. Section 3 is a review of the basis and application of BIC. Issues surrounding the assumption of a true model, the role of sample size in model selection when a true model is assumed, and real world issues such as the existence of tapering effect sizes are reviewed. Section 4 is a derivation of AIC as a Bayesian result; this derivation hinges on the use of a “savvy" prior on models. Often, model priors attempt to be noninformative; however, this practice has hidden and important implications (it is not innocent). Section 5 introduces several philosophical issues and comparisons between AIC vs. BIC. This section focuses additional attention on truth, approximating models of truth, and the careless notion of true models (mathematical models that exactly express full reality). Model selection philosophy should not be based simple Bayesian vs. non-Bayesian arguments.

Section 6 compares the performance of AIC versus BIC and notes that many Monte Carlo simulations are aimed only at assessing the probability of finding the true model. This practice misses the point of statistical inference and has lead to widespread misunderstandings. Section 6 also makes the case for multimodel inference procedures, rather than making inference from only the model estimated to be best. Multimodel inference often lessens the performance differences between AIC and BIC selection. Finally, section 7 presents a discussion of the more important issues and concludes that model selection should be viewed as a way to compute model weights (posterior model probabilities), often as a step toward model averaging and other forms of multimodel inference.

2. AIC: AN ASYMPTOTICALLY UNBIASED ESTIMATOR OF EXPECTED K-L INFORMATION

2.1 SCIENCE PHILOSOPHY AND THE INFORMATION-THEORETIC APPROACH

Information theorists do not believe in the notion of true models. Models, by definition, are only approximations to unknown reality or truth; there are no true models that perfectly reflect full reality. George Box made the famous statement “All models are wrong but some are useful." Further, a “best model," for analysis of data, depends on sample size; smaller effects can often only be revealed as sample size increases. The amount of information in large data sets (e.g., n = 3,500) greatly exceeds the information in small data sets (e.g., n = 22). Data sets in some fields are very large (terabytes) and good approximating models for such applications are often highly structured and parameterized compared to more typical applications where sample size is modest. The information-theoretic paradigm rests on the assumption that good data, relevant to the issue, are available and these have been collected in an appropriate manner (Bayesians would want this also). Three general principles guide model-based inference in the sciences.

Simplicity and Parsimony – Occam's Razor suggests “shave away all but what is necessary." Parsimony enjoys a featured place in scientific thinking in general and in modeling specifically (see Forster and Sober 1994; Forster 2000, 2001) for a strictly science philosophy perspective). Model selection (variable selection in regression is a special case) is a bias versus variance trade-off and this is the statistical principle of parsimony. Inference under models with too few parameters (variables) can be biased, while with models having too many parameters (variables) there may be poor precision or identification of effects that are, in fact, spurious. These considerations call for a balance between under- and over-fitted models – the so-called “model selection problem" (see Forster 2000, 2001).

Multiple Working Hypotheses – Chamberlin (1890, 1965) advocated the concept of “multiple working hypotheses." Here, there is no null hypothesis, instead there are several well-supported hypotheses (equivalently, “models") that are being entertained. The a priori “science" of the issue enters at this important stage. Relevant empirical data are then gathered, analyzed, and it is expected that the results tend to support one or more hypotheses, while providing less support for other hypotheses. Repetition of this general approach leads to advances in the sciences. New or more elaborate hypotheses are added, while hypotheses with little empirical support are gradually dropped from consideration. At any one point in time, there are multiple hypotheses (models) still under consideration – the model set evolves. An important feature of this multiplicity is that the number of alternative models should be kept small; the analysis of, say, hundreds or thousands of models is not justified except when prediction is the only objective, or in the most exploratory phases of an investigation. We have seen applications where more than a million models were fitted even though sample size was modest (60 to 200); we do not view such activities as reasonable. Similarly, a proper analysis must consider the science context and cannot successfully be based on “just the numbers."

Strength of Evidence – Providing quantitative information to judge the “strength of evidence" is central to science. Null hypothesis testing only provides arbitrary dichotomies (e.g.,

significant vs. nonsignificant) and in the all-to-often-seen case where the null hypothesis is false on a priori grounds the test result is superfluous. Hypothesis testing is particularly limited in model selection and this is well documented in the statistical literature. Royall (1997) provides an interesting discussion of the likelihood-based strength of evidence approach in simple statistical situations.

2.2 KULLBACK-LEIBLER INFORMATION

In 1951 S. Kullback and R. A. Leibler published a now-famous paper (Kullback and Leibler 1951) that quantified the meaning of “information" as related to R. A. Fisher's concept of sufficient statistics. Their celebrated result, called Kullback-Leibler information, is a fundamental quantity in the sciences and has earlier roots back to Boltzmann's concept of entropy (Boltzmann 1877). Boltzmann's entropy and the associated Second Law of Thermodynamics represents one of the most outstanding achievements of 19th century science.

We begin with the concept that f denotes full reality or truth; f has no parameters (parameters are a human concept). We use g to denote an approximating model, a probability distribution. Kullback-Leibler information $I(f, g)$ is the information lost when model g is used to approximate f; this is defined for continuous functions as the integral

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

Clearly the best model loses the least information relative to other models in the set; this is equivalent to minimizing $I(f, g)$, over g. Alternatively, K-L information can be conceptualized as a “distance" between full reality and a model.

Full reality f is considered to be fixed and only g varies over a space of models indexed by θ. Of course, full reality is not a function of sample size n; truth does not change as n changes. No concept of a true model is implied here and no assumption is made that the models must be nested.

The criterion $I(f, g)$ cannot be used directly in model selection because it requires knowledge of full truth, or reality, and the parameters θ in the approximating models, g_i (or, more explicitly, $g_i(x \mid \theta)$). In data analysis the model parameters must be estimated and there is often substantial uncertainty in this estimation. Models based on estimated parameters represent a major distinction from the case where model parameters are known. This distinction affects how K-L information must be used as a basis for model selection and ranking and requires a change in the model selection criterion to that of minimizing expected estimated K-L information rather than minimizing known K-L information (over the set of R models considered).

K-L information can be expressed as

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

or

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

where the expectations are taken with respect to truth. The quantity $\mathrm{E}_f[\log(f(x))]$ is a constant (say, C) across models. Hence,

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

where

$$C = \int f(x)\log(f(x))\,dx$$

does not depend on the data or the model. Thus, only relative expected K-L information, $E_f[\log(g(x \mid \theta))]$, needs to be estimated for each model in the set.

2.3 AKAIKE'S INFORMATION CRITERION, AIC

Akaike (1973, 1974, 1985, 1994) showed that the critical issue for getting a rigorous model selection criterion based on K-L information was to estimate

$$E_y E_x [\log(g(x \mid \hat{\theta}(y)))]$$

where the inner part is just $E_f[\log(g(x \mid \theta))]$ with θ replaced by the maximum likelihood estimator of θ based on the assumed model g and data y. Whereas only y denotes data, it is convenient to conceptualize both x and y as independent random samples from the same distribution. Both statistical expectations are taken with respect to truth (f). This double expectation is the target of all model selection approaches based on K-L information (e.g., AIC, AIC_c and TIC).

Akaike (1973, 1974) found a formal relationship between K-L information (a dominant paradigm in information and coding theory) and likelihood theory (the dominant paradigm in statistics) (see deLeeuw 1992). He found that the maximized log-likelihood value was a biased estimate of $E_y E_x [\log(g(x \mid \hat{\theta}(y)))]$, but this bias was approximately equal to K, the number of estimable parameters in the approximating model, g (see Burnham and Anderson 2002, chapter 7 for details). This is an asymptotic result of fundamental importance. Thus, an approximately unbiased estimator of $E_y E_x [\log(g(x \mid \hat{\theta}(y)))]$ for large samples and “good" models, is
$\log(\mathcal{L}(\hat{\theta} \mid data)) - K$. This result is equivalent to

$$\log(\mathcal{L}(\hat{\theta} \mid data)) - K = C - \hat{E}_{\hat{\theta}} [I(f, \hat{g})] ,$$

where $\hat{g} = g(\cdot \mid \hat{\theta})$.

This finding makes it possible to combine estimation (i.e., maximum likelihood or least squares) and model selection under a unified optimization framework. Akaike found an estimator of expected, relative K-L information based on the maximized log-likelihood function, corrected for asymptotic bias,

$$\text{relative } \hat{E}(\text{K-L}) = \log(\mathcal{L}(\hat{\theta} \mid data)) - K .$$

K is the asymptotic bias correction term and is in no way arbitrary (as is sometimes erroneously stated in the literature). Akaike (1973, 1974) multiplied this simple but profound result by –2 (for "historical reasons") and this became "Akaike's information criterion" (AIC),

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

In the special case of least squares (LS) estimation with normally distributed errors AIC can be expressed as

$$\text{AIC} = n\log(\hat{\sigma}^2) + 2K,$$

where

$$\hat{\sigma}^2 = \frac{\sum(\hat{\epsilon}_i)^2}{n},$$

and the $\hat{\epsilon}_i$ are the estimated residuals from the fitted model. In this case *K* must be the total number of parameters in the model, including the intercept and σ^2. Thus, AIC is easy to compute from the results of LS estimation in the case of linear models or from the results of a likelihood-based analysis in general (Edwards 1992; Azzalini 1996).

Akaike's procedures are now called information-theoretic because they are based on K-L information (see Akaike 1983, 1992, 1994; Parzen et al. 1998). It is common to find literature that seems to deal only with AIC as one of many types of criteria, without any apparent understanding that AIC is an estimate of something much more fundamental: K-L information.

Assuming a set of a priori candidate models has been defined and is well supported by the underlying science, then AIC is computed for each of the approximating models in the set (i.e., g_i, $i = 1, 2, \ldots, R$). Using AIC the models are then easily ranked from best to worst based on the empirical data at hand. This is a simple, compelling concept, based on deep theoretical foundations (i.e., entropy, K-L information, and likelihood theory). Assuming independence of the sample variates, AIC model selection has certain cross validation properties (Stone 1974, 1977).

It seems worth noting here the large sample approximate expected value of AIC (for a "good" model), in as much as this result is not given in Burnham and Anderson (2002). The MLE $\hat{\theta}(y)$ converges, as n gets large, to the θ_o that minimizes K-L information loss for model g. Large sample expected AIC converges to

$$\mathrm{E(AIC)} = -2C + 2I(f, g(\cdot \mid \theta_o)) + K .$$

2.4 IMPORTANT REFINEMENTS: EXTENDED CRITERIA

Akaike's approach allowed model selection to be firmly based on a fundamental theory and allowed further theoretical work. When K is large relative to sample size n (which includes when n is small, for any K) there is a small sample (second order bias correction) version called AIC_c,

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

(see Sugiura 1978; Hurvich and Tsai 1989, 1995), and this should be used unless $n/K >$ about 40 for the model with the largest value of K. A pervasive mistake in the model selection literature is the use of AIC when AIC_c really should be used. Because AIC_c converges to AIC, as n gets large, in practice AIC_c should be used. People often conclude that AIC overfits because they failed to use the second order criterion, AIC_c.

Takeuchi (1976) derived an asymptotically unbiased estimator of relative, expected Kullback-Leibler information that applies in general without assuming that model g is true (i.e., without the special conditions underlying Akaike's derivation of AIC). His method (TIC for Takeuchi's Information Criterion) requires quite large sample sizes to reliably estimate the bias adjustment term, which is the trace of the product of two K by K matrices (i.e., $\text{tr}[J(\theta_o)I(\theta_o)^{-1}]$, details in Burnham and Anderson 2002, pp. 65-66, 362-374). TIC represents an important conceptual advance and further justifies AIC. In many cases, the complicated bias adjustment term is approximately equal to K and this result gives further credence to using AIC and AIC_c in practice. In a sense, AIC is a parsimonious approach to TIC. The large sample expected value of TIC is $E(\text{TIC}) = -2C + 2I(f, g(\cdot \mid \theta_o)) + \text{tr}[J(\theta_o)I(\theta_o)^{-1}]$.

Investigators working in applied data analysis have several powerful methods for ranking models and making inferences from empirical data to the population or process of interest. In practice, one need not assume that the “true model" is in the set of candidates (although this is sometimes mistakenly stated in the technical literature on AIC). These information criteria are estimates of relative, expected K-L information and are an extension of Fisher's likelihood theory (Akaike 1992). AIC and AIC_c are easy to compute and quite effective in very wide variety of applications.

2.5 Δ_i VALUES

The individual AIC values are not interpretable as they contain arbitrary constants and are much affected by sample size (we have seen AIC values ranging from –600 to 340,000). Here it is imperative to rescale AIC or AIC_c to

$$\Delta_i = AIC_i - AIC_{min}$$

where AIC_{min} is the minimum of the R different AIC_i values (i.e., the minimum is at $i = min$). This transformation forces the best model to have $\Delta = 0$, while the rest of the models have positive values. The constant representing $E_f[\log(f(x))]$ is eliminated from these Δ_i values. Hence, Δ_i is the information loss experienced if we using fitted model g_i rather than the best model, g_{min}, for inference. These Δ_i allow meaningful interpretation without the unknown scaling constants and sample size issues that enter into AIC values.

The Δ_i are easy to interpret and allow a quick “strength of evidence" comparison and ranking of candidate hypotheses or models. The larger the Δ_i, the less plausible is fitted model i as being the best approximating model in the candidate set. It is generally important to know which model (hypothesis) is second best (the ranking) as well as some measure of its standing with respect to the best model. Some simple rules of thumb are often useful in assessing the relative merits of models in the set: models having $\Delta_i \leq 2$ have substantial support (evidence), those where $4 \leq \Delta_i \leq 7$ have considerably less support, while models having $\Delta_i > 10$ have essentially no support. These rough guidelines have similar counterparts in the Bayesian literature (Raftery 1996).

Naive users often question the importance of a $\Delta_i = 10$ when the two AIC values might be, for example, 280,000 and 280,010. The difference of 10 here might seem trivial. In fact, large AIC values contain large scaling constants, while the Δ_i are free of such constants. Only these differences in AIC are interpretable as to strength of evidence.

2.6 LIKELIHOOD OF A MODEL GIVEN THE DATA

The simple transformation $\exp(-\Delta_i/2)$, for $i = 1, 2, \ldots, R$, provides the likelihood of the model (Akaike 1981) given the data: $\mathcal{L}(g_i \mid data)$. [Recall, Akaike defined his AIC after multiplying through by –2; otherwise, $\mathcal{L}(g_i \mid data) = \exp(\Delta_i)$ would have been the case, with Δ redefined in the obvious way]. This is a likelihood function over the model set in the sense that $\mathcal{L}(\theta \mid data, g_i)$ is the likelihood over the parameter space (for model g_i) of the parameter θ given the data (x) and the model (g_i).

The relative likelihood of model i versus model j is $\mathcal{L}(g_i \mid data)/\mathcal{L}(g_j \mid data)$; this is termed the evidence ratio and it does not depend on any of the other models under consideration. Without loss of generality we may assume model g_i is more likely than g_j. Then if this evidence ratio is large (e.g., > 150 is quite large), model g_j is a poor model relative to model g_i, based on the data.

2.7 AKAIKE WEIGHTS, w_i

It is convenient to normalize the model likelihoods such that they sum to 1 and treat them as probabilities, hence we use

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

The w_i, called Akaike weights, are useful as the “weight of evidence" in favor of model $g_i(\cdot \mid \theta)$ as being the actual K-L best model in the set (in this context a model, g, is considered as a “parameter"). The ratios w_i/w_j are identical to the original likelihood ratios, $\mathcal{L}(g_i \mid data)/\mathcal{L}(g_j \mid data)$, so are invariant to the model set, but the w_i values depend on the full model set because the sum to 1. However, w_i, $i = 1, \ldots, R$ are useful in additional ways. For example, the w_i are interpreted as the probability that model i is, in fact, the K-L best

model for the data (strictly under K-L information theory this is an heuristic interpretation, but it is justified by a Bayesian interpretation of AIC – see below). This latter inference about model selection uncertainty is conditional on both the data and the full set of a priori models considered.

2.8 UNCONDITIONAL ESTIMATES OF PRECISION, A TYPE OF MULTIMODEL INFERENCE

Typically, estimates of sampling variance are conditional on a given model as if there was no uncertainty about which model to use (Breiman called this a “quiet scandal," Breiman 1992). When model selection has been done, there is a variance component due to model selection uncertainty that should be incorporated into estimates of precision. That is, one needs estimates that are “unconditional" on the selected model. A simple estimator of the unconditional variance for the maximum likelihood estimator $\hat{\theta}$ from the selected (best) model is,

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

where,

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

and $\bar{\theta}$ represents a form of “model averaging." The notation $\hat{\theta}_i$ here means the parameter θ is estimated based on model g_i, but θ is a parameter in common to all R models (even if its value is 0 in model k, so that then we use $\hat{\theta}_k = 0$). This estimator, from Buckland et al. (1997), includes a term for the conditional sampling variance, given model g_i (denoted as $\widehat{\text{var}}(\hat{\theta}_i \mid g_i)$ here) and a variance component for model selection uncertainty, $(\hat{\theta}_i - \bar{\theta})^2$. These variance

components are multiplied by the Akaike weights, which reflect the relative support, or evidence, for model *i*. Burnham and Anderson (2002, pp. 206-243) provide a number of Monte Carlo results on achieved confidence interval coverage when information-theoretic approaches are used in some moderately challenging data sets. For the most part, achieved confidence interval coverage is near the nominal level. Model averaging arises naturally when the unconditional variance is derived.

2.9 OTHER FORMS OF MULTIMODEL INFERENCE

Rather than base inferences on a single, selected best model from an a priori set of models, inference can be based on the entire set of models. Such inferences can be made if a parameter, say θ, is in common over all models (as θ_i in model g_i), or if the goal is prediction. Then by using the weighted average for that parameter across models (i.e., $\bar{\hat{\theta}} = \sum w_i \hat{\theta}_i$) we are basing point inference on the entire set of models. This approach has both practical and philosophical advantages. Where a model averaged estimator can be used it often has a more honest measure of precision and reduced bias compared to the estimator from just the selected best model (Burnham and Anderson 2002, chapters 4–6). In all-subsets regression we can consider the regression coefficient (parameter) β_p for predictor x_p is in all the models, but for some models $\beta_p = 0$ (x_p is not in those models). In this situation if model averaging is done over all the models the resultant estimator $\tilde{\beta}_p$ has less model selection bias than $\hat{\beta}_p$ taken from the selected best model (Burnham and Anderson 2002, pp. 151-153, 248-255; Lukacs, in review).

Assessment of the relative importance of variables has often been based only on the best model (e.g., often selected using a stepwise testing procedure). Variables in that best model are considered “important" while excluded variables are considered not important. This is too simplistic. Importance of a variable can be refined by making inference from all the models in the candidate set (see Burnham and Anderson 2002, chapters 4–6). Akaike weights are summed for all models containing predictor variable x_j, $j = 1, \ldots, R$; denote

these sums as $w_+(j)$. The predictor variable with the largest predictor weight, $w_+(j)$, is estimated to be the most important; the variable with the smallest sum is estimated to be the least important predictor. This procedure is superior to making inferences concerning the relative importance of variables based only on the best model. This is particularly important when the second or third best model is nearly as well supported as the best model, or when all models have nearly equal support. (There are “design" considerations about the set of models to consider when a goal is assessing variable importance, we do not discuss these considerations here – the key issue is one of balance of models with and without each variable).

2.10 SUMMARY

At a conceptual level, reasonable data and a good model allow a separation of “information" and “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 want an approximating model that minimizes information loss, $I(f, g)$, and properly separates noise (non-information or entropy) from structural information. In a very important sense, we are not trying to model the data; instead, we are trying to model the information in the data.

Information-theoretic methods are relatively simple to understand and practical to employ across a very large class of empirical situations and scientific disciplines. The methods are easy to compute by hand if necessary, assuming one has the parameter estimates, the conditional variances $\widehat{\text{var}}(\hat{\theta}_i \mid g_i)$, and the maximized log-likelihood values for each of the R candidate models from standard statistical software. Researchers can easily understand the heuristics and application of the information-theoretic methods; we believe it is *very* important that people understand the methods they employ. Information-theoretic approaches should not be used unthinkingly; a good set of a priori models is essential and this involves professional judgment and integration of the science of the issue into the model set.

3. UNDERSTANDING BIC

Schwarz (1978) derived the Bayesian information criterion as

$$\text{BIC} = -2\ln(\mathcal{L}) + K\log(n) .$$

As usually used one computes BIC for each model and selects the model with the smallest criterion value. BIC is a misnomer as it is not related to information theory. As with ΔAIC_i we define ΔBIC_i as the difference of BIC for model g_i and the minimum BIC value. More complete usage entails computing posterior model probabilities, p_i, as

$$p_i = \Pr\{g_i \mid data\} = \frac{\exp(-\frac{1}{2}\Delta\text{BIC}_i)}{\sum_{r=1}^{R}\exp(-\frac{1}{2}\Delta\text{BIC}_r)}$$

(Raftery 1995). The above posterior model probabilities are based on assuming prior model probabilities are all $1/R$. Most applications of BIC use it in a frequentist spirit, hence ignore issues of prior and posterior model probabilities.

The model selection literature, as a whole, is confusing as regards the following issues about BIC (and about Bayesian model selection in general):

1) Does the derivation of BIC assume the existence of a true model; or more narrowly, is the true model assumed to be in the model set when using BIC? (Schwarz's derivation specified these conditions.)
2) What do the “model probabilities" mean; that is, how should we interpret them vis-a-vis a “true" model?

Mathematically (we emphasize “mathematical" here), for an iid sample and a fixed set of models, there is a model, say model g_t , with posterior probability p_t such that as $n \to \infty$ then $p_t \to 1$ and all other $p_r \to 0$. In this sense there is a clear target model that BIC “seeks" to select.

3) Does the above result mean model g_t must be the true model?

The answers to questions 1 and 3 are simple: No. That is, BIC (as the basis for an approximation to a certain Bayesian integral) can be derived without assuming the model underlying the derivation is true (see e.g., Cavanaugh and Neath 1999; Burnham and Anderson 2002, pp. 293-295). Certainly in applying BIC the model set need not contain the (nonexistent) true model representing full reality. Moreover, the convergence in probability of the BIC selected model to a target model (under the idealization of an iid sample) does not logically mean that that target model must be the true data generating distribution.

The answer to question 2 involves characterizing the target model to which the BIC-selected model converges. That model can be characterized in terms of the values of the Kullback-Leibler discrepancy and K for the set of models. For model g_r the Kullback-Leibler "distance" of the model from truth is denoted $I(f, g_r)$. Often, $g_r \equiv g_r(x \mid \theta)$ would denote a parametric family of models for $\theta \in \Theta$, Θ being a K_r dimensional space. However, we take g_r generally to denote the specific family member for the unique $\theta_o \in \Theta$ which makes g_r, in the family of models, closest to truth in K-L distance. For the family of models $g_r(x \mid \theta)$, $\theta \in \Theta$, as $n \to \infty$ (with iid data) the MLE and Bayesian point estimator of θ converge to θ_o. Thus asymptotically we can characterize the particular model that g_r represents: $g_r \equiv g_r(x \mid \theta_o)$ (for details see Burnham and Anderson 2002 and references cited therein). Also, we have the set of corresponding minimized K-L distances: $\{I(f, g_r), r = 1, \ldots, R\}$. For an iid sample we can represent these distances as $I(f, g_r) = nI_1(f, g_r)$ where the $I_1(f, g_r)$ do not depend on sample size (they are for $n = 1$). The point of this representation is to emphasize than the effect of increasing sample size is to scale-up these distances.

We may assume, without loss of generality, that these models are indexed worst (g_1) to best (g_R) in terms of their K-L distance and dimension K_r, hence $I(f, g_1) \geq I(f, g_2) \geq \cdots \geq I(f, g_R)$. Figures 1-3 show three hypothetical scenarios of how these ordered distances might appear for $R = 12$ models, for unspecified n (since n serves merely to scale the y-axis). Let Q be the tail-end subset, of the so-ordered models, defined by $\{g_r, r \geq t, 1 \leq t \leq R \mid I(f, g_{t-1}) > I(f, g_t) = \cdots = I(f, g_R)\}$. Set Q exists because $t = R$ (and

$t = 1$) is allowed, in which case the K-L best model (of the R models) is unique. For the case when subset Q contains more than one model (i.e., $1 \leq t < R$) then all of the models in this subset have the same K-L distance. Therefore, we further assume that models g_t to g_R are ordered such that $K_t < K_{t+1} \leq \cdots \leq K_R$ (in principle $K_t = K_{t+1}$ could occur).

Thus, model g_t is the most parsimonious model of the subset of models that are tied for K-L best model. In this scenario (iid sample, fixed model set, $n \to \infty$) the BIC-selected model converges with probability 1 to model g_t and p_t converges to 1. However, unless $I(f, g_t) = 0$ model g_t is not identical to f (nominally considered as truth), so we call it a quasi-true model. The only truth here is that in this model set, models g_{t+1} to g_R provide no improvement over model g_t – they are unnecessarily general (independent of sample size). The quasi-true model in the set of R models is the most parsimonious model that is closest to truth in K-L information loss (model 12 in Figures 1 and 3, model 4 in Figure 2).

Thus, the Bayesian posterior model probability p_r is the inferred probability that model g_r is the quasi-true model in the model set. For a “very large" sample size model g_t is the best model to use for inference. However, for small or moderate sample sizes obtained in practice the model selected by BIC may be much more parsimonious than model g_t, especially if the quasi-true model is the most general model, g_R as in Figure 1. The concern for realistic sample sizes then is that the BIC-selected model may be underfit at the given n. The model selected by BIC approaches the BIC target model from below, as n increases, in terms of the ordering we imposed on the model set. This selected model can be quite far from the BIC theoretical target model at sample sizes seen in practice when tapering effects are present (Figure 1). The situation where BIC performs well is that of Figure 2 with suitably large n).

figures 1 and 2 about here

Moreover, the BIC target model does not depend on sample size n. However, we know that the number of parameters we can expect to reliably estimate from finite data does depend on n. In particular, if the set of ordered (large to small) K-L distances show tapering effects (Figure 1) then a best model for making inference from the data may well be a more

parsimonious model than the BIC target model (g_{12} in Figure 1), such as the best expected estimated K-L model, which is the AIC target model. As noted above the target model for AIC is the model that minimizes $\mathrm{E}_f\,[I(f, g_r(\cdot \mid \hat{\theta}))]$, $r = 1, \ldots, R$. This target model is specific for the sample size at hand, hence AIC seeks a best model as its target, where best is heuristically a bias-variance trade-off (not a quasi-true model).

In reality one can only assert that BIC model selection is asymptotically consistent for the (generally) unique quasi-true model in the set of models. But that BIC-selected model can be quite biased at not-large n as an estimator of its target model. Also, from an inference point of view observing p_t is nearly 1 does not justify an inference that model g_t is truth (such a statistical inference requires an a-prior certainty that the true model is in the model set). This issue is intimately related to the fact that only differences such as $I(f, g_r) - I(f, g_t)$ are estimable from data (these K-L differences are closely related to $\mathrm{AIC}_r - \mathrm{AIC}_t$ differences, hence to the Δ). Hence, with model selection the effect is that sometimes people are erroneously lulled into thinking (assuming) $I(f, g_t)$ is 0 and hence thinking they have found (the model for) full reality. These fitted models sometimes have 7 or fewer parameters; surely full reality cannot be so simple in the life sciences, economics, medicine and the social sciences.

4. AIC AS A BAYESIAN RESULT

BIC model selection arises in the context of a large sample approximation to the Bayes factor conjoined with assuming equal priors on models. The BIC statistic can be used more generally with any set of model priors. Let q_i be the prior probability placed on model g_i. Then the Bayesian posterior model probability is approximated as

$$\Pr\{g_i \mid data\} = \frac{\exp(-\frac{1}{2}\Delta\mathrm{BIC}_i)q_i}{\sum_{r=1}^{R}\exp(-\frac{1}{2}\Delta\mathrm{BIC}_r)q_r}$$

(this posterior actually depends on not just the data but also upon the model set and the prior distribution on those models). Akaike weights can be easily obtained by using the model prior q_i as proportional to

$$\exp(\tfrac{1}{2}\Delta\mathrm{BIC}_i) \cdot \exp(-\tfrac{1}{2}\Delta\mathrm{AIC}_i) \, .$$

Clearly,

$$\exp(-\tfrac{1}{2}\Delta\mathrm{BIC}_i) \cdot \exp(\tfrac{1}{2}\Delta\mathrm{BIC}_i) \cdot \exp(-\tfrac{1}{2}\Delta\mathrm{AIC}_i) = \exp(-\tfrac{1}{2}\Delta\mathrm{AIC}_i) \, ;$$

hence, with the implied prior probability distribution on models we get

$$p_i = \Pr\{g_i \mid data\} = \frac{\exp(-\frac{1}{2}\Delta\mathrm{BIC}_i)q_i}{\sum_{r=1}^{R}\exp(-\frac{1}{2}\Delta\mathrm{BIC}_r)q_r} = \frac{\exp(-\frac{1}{2}\Delta\mathrm{AIC}_i)}{\sum_{r=1}^{R}\exp(-\frac{1}{2}\Delta\mathrm{AIC}_r)} = w_i \, ,$$

which is the Akaike weight for model g_i.

This prior probability on models can be expressed in a simple form as

$$q_i = C \cdot \exp(\tfrac{1}{2}K_i \log(n) - K_i) \tag{3a}$$

where

$$C = \frac{1}{\sum_{r=1}^{R}\exp(\frac{1}{2}K_r \log(n) - K_r)} \, . \tag{3b}$$

Thus, formally, the Akaike weights from AIC are (for large samples) Bayesian posterior model probabilities for this model prior (more details are in Burnham and Anderson 2002, pp. 302-305).

Given a model $g(x \mid \theta)$ the prior distribution on θ will not, should not, depend on sample size. This is very reasonable. Probably following from this line of reasoning, traditional Bayesian thinking about the prior distribution on models has been that q_r, $r = 1$,

$\ldots$, R, would also not depend on n or K_r. This approach is neither necessary, nor reasonable. There is limited information in a sample so the more parameters one estimates, the poorer the average precision becomes for these estimates. Hence, in considering the prior distribution q on models we must consider the context of what we are assuming about the information in the data, as regards parameter estimation, and the models as approximations to some conceptual underlying "full truth" generating distribution. While $q_r = 1/R$ seems reasonable and innocent, it is not always reasonable and is never innocent: i.e., it implies the target model is truth rather than a best approximating model given parameters are to be estimated. This is an important and unexpected result.

It is useful to think in terms of effects, for individual parameters, as $|\theta|/\text{se}(\hat{\theta})$. The standard error depends on sample size, hence effect-size depends on sample size. We would assume for such effects that few or none are truly zero in the context of analysis of real data from complex observational, quasi-experimental or experimental studies (i.e., Figure 1 applies). In the information-theoretic spirit we assume meaningful, informative data and thoughtfully selected predictors and models (not all studies meet these ideals). We assume tapering effects: some may be big (values like 10 or 5), but some are only 2, 1 or 0.5, or less. We assume we can only estimate n/m parameters reliably; m might be 20 or as small as 10 (but surely $m \gg 1$ and $m \ll 100$). (In contrast, the scenario where BIC performs better than AIC is one where it is assumed there are a few big effects defining the quasi-true model which is itself nested in several, or many, overly general models, i.e. Figure 2 applies).

These concepts imply that the size (i.e., K) of the appropriate model to fit to data should logically depend on n. This idea is not foreign to the statistical literature. For example, Lehman (1990, p. 160) attributes to R. A. Fisher the quote "More or less elaborate forms will be suitable according to the volume of the data." Using the notation k_0 for the optimal K, Lehman (1990, p. 162) goes on the say "The value of k_0 will tend to increase as the number of observations increases and its determination thus constitutes implementation of Fisher's

suggestion … ." From a recent book (Williams 2001, p. 235), and quoted exactly, "… *we CANNOT ignore the degree of resolution of the experiment when choosing our prior*."

These ideas lead to a model prior wherein conceptually q_r should depend on n and K_r. Such a prior (class of priors, actually) we call a "savvy prior." A savvy (definition: shrewdly informed) prior is logical under the information-theoretic model selection paradigm. We will call the savvy prior on models given by

$$q_i = C \cdot \exp(\tfrac{1}{2} K_i \log(n) - K_i)$$

(formula 3b gives C) the K-L model prior. It is unique in terms of producing AIC as approximately a Bayesian procedure (approximate only because BIC is an approximation).

Alternative savvy priors might be based on distributions such as a modified Poisson (i.e., applied to only K_r, $r = 1, \ldots, R$) with expected K set to be $n/10$. We looked at this idea in an all subsets selection context and found that the K-L model prior produces a more spread-out (higher entropy) prior as compared to such a Poisson-based savvy prior when both produced the same E(K). We are not wanting to start a cottage industry of seeking a best savvy prior because model averaged inference seems very robust to model weights when those weights are well founded (as is the case for Akaike weights).

The full implications of being able to interpret AIC as a Bayesian result have not been determined and is an issue outside the scope of this paper. It is, however, worth mentioning that the model-averaged Bayesian posterior is a mixture distribution of each model-specific posterior distribution, with weights being the posterior model probabilities. Therefore, for any model averaged parameter estimator, and in particular for model averaged prediction, alternative variance and covariance formulae are

$$\widehat{\text{var}}(\bar{\hat{\theta}}) = \sum_{i=1}^{R} w_i \left[\widehat{\text{var}}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \bar{\hat{\theta}})^2 \right], \qquad (4)$$

$$\widehat{\text{cov}}(\hat{\bar{\theta}}, \hat{\bar{\tau}}) = \sum_{i=1}^{R} w_i \left[\widehat{\text{cov}}(\hat{\theta}_i, \hat{\tau}_i \mid g_i) + (\hat{\theta}_i - \hat{\bar{\theta}})(\hat{\tau}_i - \hat{\bar{\tau}}) \right] . \quad (5)$$

The formula given in Burnham and Anderson (2002, pp 163-164) for such an unconditional covariance is ad hoc; hence we now recommend the above covariance formula. We have re-run many simulations and examples from Burnham and Anderson (1998) using variance formula (4) and found its performance is almost identical to that of the original unconditional variance formula (1) (see also Burnham and Anderson 2002, pp. 344-345). Our pragmatic thought is that it may well be desirable to use formula (4) rather than (1), but it is not necessary, except when covariances (formula 5) are also computed.

5. RATIONAL CHOICE OF AIC OR BIC

5.1 FREQUENTIST VERSUS BAYESIAN IS NOT THE ISSUE

The model selection literature contains, defacto, a long running debate about using AIC or BIC. Much of the purely mathematical or Bayesian literature recommends BIC. We maintain that almost all the arguments for use of BIC rather than AIC, with real data, are flawed and hence they contribute more to confusion than to understanding. This assertion by itself is not an argument for AIC or against BIC because there are clearly defined contexts where each method out performs the other (Figures 1 or 2 for AIC or BIC, respectively).

For some people BIC is strongly preferred because it is a Bayesian procedure, and they think AIC is non-Bayesian. However, AIC model selection is just as much a Bayesian procedure as is BIC selection. The difference is in the prior distribution placed on the model set. Hence, for a Bayesian the argument about BIC versus AIC must reduce to one about priors on the models.

Alternatively, both AIC and BIC can be argued for, or derived, under a non-Bayesian approach. We have given above the arguments for AIC. When BIC is so derived it is usually motivated by the mathematical context of nested models including a true model simpler than

the most general model in the set. This corresponds to the context of Figure 2, except with the added (but not needed) assumption that $I(f, g_t) = 0$. Moreover, the goal is taken to be selection of this true model with probability 1 as $n \to \infty$ (asymptotic consistency; or sometimes dimension consistency).

Given that AIC and BIC model selection can both be derived as either frequentist or Bayesian procedures one cannot argue for or against either of them on the basis that it is, or is not, Bayesian or non-Bayesian. What fundamentally distinguishes AIC versus BIC model selection is their different philosophies, including the exact nature of their target models, and the conditions under which one outperforms the other for performance measures such as predictive mean square error. Thus we maintain that comparison, hence selection for use, of AIC versus BIC must be based on comparing measures of their performance under conditions realistic of applications. (A, now rare, version of Bayesian philosophy would deny the validity of such hypothetical frequentist comparisons as a basis for justifying inference methodology. We regard such nihilism as being outside of the evidential spirit of science; we demand evidence).

5.2 DIFFERENT PHILOSOPHIES AND TARGET MODELS.

We have given the different philosophies and contexts in which the AIC or BIC model-selection criteria arise and can be expected to perform well. Here we explicitly contrast these underpinnings in terms of K-L distances for the model set $\{g_r(x \mid \theta_o), r = 1, \ldots, R\}$ with reference to Figures 1, 2, and 3, which represent $I(f, g_r) = nI_1(f, g_r)$. Sample size n is left unspecified except it is large relative to K_R, the largest value of K_r, yet of a practical size (e.g., $K_R = 15$ and $n = 200$).

Given that the model parameters must be estimated so that parsimony is an important consideration then just by looking at Figure 1 we cannot say what is the best model to use for inference as a model fitted to the data. Model 12, as $g_{12}(x \mid \theta_o)$ (i.e., at θ being the K-L-distance minimizing parameter value in Θ for this class of models) is the best theoretical

model, but $g_{12}(x \mid \hat{\theta})$ may not be the best model for inference. Model 12 is the target model for BIC, but not for AIC. The target model for AIC will depend on *n* and could be, for example, model 7 (there would be an *n* for which this would be true).

Despite that the target of BIC is a more general model than the target model for AIC, the model most often selected here by BIC will be less general than model 7 unless *n* is very large. It might be model 5 or 6. It is known (numerous papers and simulations in the literature) that in the tapering effects context (Figure 1) AIC performs better than BIC. If this is the context of one's real data analysis, they should use AIC.

A very different scenario is given by Figure 2, wherein there are a few big effects, all captured by model 4 (i.e., $g_4(x \mid \theta_o)$), and models 5 to 12 do not improve at all on model 4. This scenario generally corresponds model 4 being nested in models 5 to 12, often as part of a full sequence of nested models, $g_i \subset g_{i+1}$. The obvious target model for selection is model 4; models 1 to 3 are too restrictive and models in the class of models 5 to 12 contain unneeded parameters (such parameters are actually 0). Scenarios like that of Figure 2 are often used in simulation evaluations of model selection, despite that they seem unrealistic for most real data, so conclusions do not logically extend to the Figure 1 (or Figure 3) scenario.

Under the Figure 2 scenario and for sufficiently large *n*, BIC often selects model 4 and does not select more general models (but may select less general models). AIC will select model 4 much of the time, will tend not to select less general models, but will sometimes select more general models and do so even if *n* is large. It is this scenario that motivates the model selection literature to conclude BIC is consistent and AIC is not consistent. We maintain that this conclusion is for an unrealistic scenario with respect to a lot of real data as regards the pattern of the K-L distances. Also ignored in this conclusion is the issue that for real data the model set itself should change as sample size increases by orders of magnitude. Also, inferentially such “consistency" can only imply a quasi-true model, not truth as such.

That reality could be as depicted in Figure 2 seems strained, but it could be as depicted in Figure 3 (as well as Figure 1). The latter scenario might occur and presents a problematic

case for theoretical analysis. Simulation seems needed there, and in general to evaluate model selection performance under realistic scenarios. For Figure 3 the target model for BIC is also model 12, but model 5 would likely be a better choice at moderate to even large sample sizes.

figure 3 about here

5.3 FULL REALITY AND TAPERING EFFECTS

Often the context of data analysis with a focus on model selection is one of many covariates and predictive factors (x). The conceptual truth underlying the data is about what is the marginal truth just for this context and these measured factors. If this truth, conceptually as $f(y \mid x)$, implies $\mathrm{E}(y \mid x)$ has tapering effects then any fitted good model will need tapering effects. In the context of a linear model, and for an unknown (to us) ordering of the predictors, then for $\mathrm{E}(y \mid x) = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$ our models will have $\mid \beta_1/\mathrm{se}(\hat{\beta}_1) \mid > \mid \beta_2/\mathrm{se}(\hat{\beta}_2) \mid > \cdots > \mid \beta_p/\mathrm{se}(\hat{\beta}_p) \mid > 0$ (β here is the K-L best parameter value given truth f and model g). It is possible that $\mid \beta_p/\mathrm{se}(\hat{\beta}_p) \mid$ would be

very small (almost 0) relative to $|\beta_1/\text{se}(\hat{\beta}_1)|$. For nested models, appropriately ordered, such tapering effects would lead to graphs like Figures 1 or 3 for either the K-L values or the actual $|\beta_r/\text{se}(\hat{\beta}_r)|$.

Whereas tapering effects for full reality are expected to require tapering effects in models and hence a context where AIC selection is called for, in principle full reality could be simple, in some sense, and yet our model set might require tapering effects. The effects (tapering or not) that matter as regards whether AIC (Figure 1) or BIC (Figure 2) model selection is the method of choice are the K-L values $I(f, g_r(\cdot \mid \beta_o))$, $r = 1, \ldots, R$, not what is implicit in truth itself. Thus, if the type of models g in our model set are a poor approximation to truth f, we can expect tapering effects for the corresponding K-L values. For example, consider the target model $\text{E}(y \mid x) = 17 + (0.3(x_1 x_2)^{0.5}) + \exp(-0.5(x_3(x_4)^2))$. However, if our candidate models are all linear in the predictors (with main effects, interactions, quadratic effects, and so forth) we will have tapering effects in the model set and AIC is the method of choice. Our conclusion is that we nearly always face some tapering effect sizes; these are revealed as sample size increases.

6. ON PERFORMANCE COMPARISONS OF AIC AND BIC

There is now ample and diverse of theory for AIC and BIC based model selection and multimodel inference, such as model averaging (as opposed to the traditional “use only the selected best model for inference"). Also, it is clear that there are different conditions under which AIC or BIC should outperform the other one in measures such as estimated mean square error. Moreover, performance evaluations and comparisons should be for actual sample sizes seen in practice, not just asymptotically; partly this is because if sample size increased substantially we should then consider revising the model set.

There are many simulation studies in the statistical literature on either AIC, or BIC alone, or often comparing them and making recommendations on which one to use. Overall, these studies have lead to confusion because either they often failed to be clear on the conditions and objectives of the simulations, or they generalized (extrapolated, actually) their conclusions beyond the specific conditions of the study. For example, were the study conditions only Figure 2 scenarios (all too often, yes) so BIC was favored? Were Figures 1, 2 and 3 scenarios all used but the author's objective was to select the true model, which was placed in the model set (and usually was a simple model), hence results were confusing and often disappointing. We submit that many reported studies are not appropriate as a basis for inference about which criterion should be used for model selection with real data.

Also, many studies, even now, only examine operating properties (e.g., confidence interval coverage and mean square error) of inference based on the use of just the selected best model (e.g., Meyer and Laud 2002). There is a strong need to evaluate operating properties of multimodel inference in scenarios realistic of real data analysis. Authors need to be very clear about the simulation scenarios used vis-a-vis the generating model; is it simple or complex, is it in the model set, and are there tapering effects. One must also be careful to note if the objective of the study was to select the true model or if it was to select a best model, as for prediction. These factors and considerations affect the conclusions from simulation evaluations of model selection. Authors should avoid sweeping conclusions based on limited, perhaps unrealistic, simulation scenarios; this error is common in the literature. Finally, to have realistic objectives the inference goal ought to be that of obtaining best predictive inference or best inference about a parameter in common to all models, rather than “select the true model."

6.1 MODEL AVERAGED VERSUS BEST MODEL INFERENCE

When prediction is the goal one can use model averaged inference rather than prediction based on a single selected best model (hereafter referred to as “best").

It is clear from the literature that has evaluated, or even considered, model averaged inferences compared to the best model strategy that model averaging is superior (e.g., Buckland et al. 1997; Hoeting et al. 1999; Wasserman 2000; Breiman 2001; Burnham and Anderson 2002; Hansen and Kooperberg 2002; Lukacs et al. in review). The method known as boosting is a type of model averaging (Hand and Vinviotti 2003, p. 130; this paper is also useful reading for its comments on truth and models). However, model averaged inference is not common, nor has there been much effort to evaluate it even in major publications on model selection nor in simulation studies on model selection; such studies all too often look only at the best model strategy. Model averaging and multimodel inference in general are deserving of more research.

As an example of predictive performance we report here some results of simulation based on the real data used in Johnson (1996). These data were originally taken to explore multiple regression to predict percent body fat based on 13 predictors (body measurements) that are easily measured. We choose these data as a focus because they were used by Hoeting et al. (1999) in illustrating BIC and Bayesian model averaging (see also Burnham and Anderson 2002, pp. 268-284). The data are from a sample of 252 males, ages 21 to 81 and are available on the web in conjunction with Johnson (1996). The web site states “The data were generously supplied by Dr. A. Garth Fisher, Human Performance Research Center, Brigham Young University, Provo, Utah 84602, who gave permission to freely distribute the data and use them for non-commercial purposes."

We take the response variable as $y = 1/D$; D is measured body density (observed minimum and maximum are 0.9950 and 1.1089). The correlations among the 13 predictors are strong, but not extreme, almost entirely positive, and range from -0.245 (age & height) to 0.941 (weight & hip circumference). The design matrix is full rank. The literature (e.g., Hoeting et al. 1999) supports that the measurements y and $\underline{x} = (x_1, \ldots, x_{13})'$ on a subject can be suitably modeled as multivariate normal. Hence, we base simulation on a joint multivariate model mimicking these 14 variables by using the observed variance-covariance matrix as

truth. From that full 14×14 observed variance-covariance matrix for y and $\underline{x}$, and theory of multivariate normal distributions, we computed for the full linear model of y regressed on $\underline{x}$ the theoretical regression coefficients and their standard errors. The resultant theoretical effect sizes $\beta_i/\text{se}(\hat{\beta}_i)$ taken as underlying the simulation are given in Table 1, ordered from largest to smallest by their absolute values. Also shown is the index (j) of the actual predictor variable as ordered in Johnson (1996).

Table 1 about here

We generated data from 13 models that range from having only one huge effect-size (generating model 1) to the full tapering effects model (#13). This was done by first generating a value of $\underline{x}$ from its assumed 13 dimensional "marginal" multivariate distribution. Then we generate $y = \text{E}(y \mid \underline{x}) + \epsilon$ (ϵ was independent of $\underline{x}$) for 13 specific models of $\text{E}_i(y \mid \underline{x})$ with $\epsilon \sim \text{normal}(0, \sigma_i^2)$, $i = 1, \ldots, 13$. Given the generating structural model on expected y, σ_i was specified so that the total expected variation (structural plus residual) in y was always the same and was equal to the total variation of y in the original data. Thus, $\sigma_1, \ldots, \sigma_{13}$ are monotonically decreasing.

For the structural data generating models we used $\text{E}_1(y \mid \underline{x}) = \beta_0 + \beta_6 x_6$ (generating model 1), $\text{E}_2(y \mid \underline{x}) = \beta_0 + \beta_6 x_6 + \beta_{13} x_{13}$ (generating model 2), and so forth. Without loss of generality we used $\beta_0 = 0$. Thus from Table 1, one can perceive the structure of each generating model reported on in Table 2. Theory asserts that under generating model 1 BIC is relatively more preferred (leads to better predictions) but as the sequence of generating models progresses K-L based model selection becomes increasingly more preferred.

Independently from each generating model we generated 10,000 samples of $\underline{x}$ and y, each of size $n = 252$. For each such sample all possible 8,192 models were fit, i.e., all subsets model selection was used based on all 13 predictor variables (regardless of the data generating model). Model selection was then applied to this set of models using both AIC_c and BIC to find the corresponding sets of model weights (posterior model probabilities), hence also the best model (with $n = 252$ and maximum K being 15 AIC_c rather than AIC should be used).

The full set of simulations took about two months of CPU time on a 1.9 GHz Pentium 4 computer.

The inference goal in this simulation was prediction. Therefore, after model fitting for each sample we generated, from the same generating model *i*, one additional statistically independent value of $\underline{x}$ and then of $E(y) \equiv E_i(y \mid \underline{x})$. Based on the fitted models from the generated sample data and this new $\underline{x}$, $E(y \mid \underline{x})$ was predicted (hence, $\hat{E}(y)$), either from the selected best model, or as the model averaged prediction. The measure of prediction performance used was predictive mean square error (PMSE) as given by the estimated (from 10,000 trials) expected value of $(\hat{E}(y) - E_i(y \mid \underline{x}))^2$.

Thus, we obtained four PMSE values from each set of 10,000 trials: PMSE for both the "best" and "model averaged" strategies for both AIC_c and BIC. Denote these as PMSE(AIC_c, best), PMSE(AIC_c, ma), PMSE(BIC, best), PMSE(BIC, ma), respectively. Absolute values of these PMSEs are not of interest here because our goal is comparison of methods; hence, in Table 2 we report only ratios of these PMSEs. The first two columns of Table 2 compare results for AIC_c to those for BIC based on the
ratios

$$\frac{\text{PMSE}(AIC_c, \text{best})}{\text{PMSE(BIC, best)}}, \quad \text{column 1, Table 2}$$

$$\frac{\text{PMSE}(AIC_c, \text{ma})}{\text{PMSE(BIC, ma)}}, \quad \text{column 2, Table 2.}$$

Thus if AIC_c produced better prediction results for generating model *i* the value in that row for columns 1 or 2 is < 1, otherwise BIC was better.

table 2 about here

The results are as qualitatively expected: under a Figure 2 scenario with only a few big effects (or no effects), such as for generating models 1 or 2, BIC outperforms AIC_c. But as we move more into a tapering effects scenario (Figure 1) AIC_c is better. We also see from Table

2, by comparing columns 1 and 2, that the performance difference of AIC_c versus BIC is reduced under model averaging: column 2 values are generally closer to 1 than are column 1 values, under the same generating model.

Columns 3 and 4 of Table 2 compare the model averaged to best model strategy within AIC_c or BIC methods:

$$\frac{\text{PMSE}(AIC_c, \text{ma})}{\text{PMSE}(AIC_c, \text{best})}, \quad \text{column 3, Table 2}$$

$$\frac{\text{PMSE}(\text{BIC}, \text{ma})}{\text{PMSE}(\text{BIC}, \text{best})}, \quad \text{column 4, Table 2.}$$

Thus if model averaged prediction is more accurate than best model prediction the value in columns 3 or 4 is < 1, which it always is. It is clear that here, for prediction, model averaging is always better than the best model strategy. The literature and our own other research on this issue suggests that such a conclusion will hold generally. A final comment about information in Table 2, columns 3 and 4: the smaller the ratio, the more beneficial was the model averaging strategy compared to the best model strategy.

In summary, we maintain that the proper way to compare AIC- and BIC-based model selection is in terms of achieved performance, especially prediction but also confidence interval coverage. In so doing it must be realized that these two criteria for computing model weights have their optimal performance under different conditions: AIC for tapering effects, BIC for when there are either no effects at all, or there are a few big effects and all others are zero effects (no intermediate effects, no tapering effects). Moreover, the extant evidence strongly supports that model averaging (where applicable) produces better performance for either AIC or BIC under all circumstances.

6.2 GOODNESS-OF-FIT AFTER MODEL SELECTION

Goodness-of-fit theory about the selected best model is a subject that has been almost totally ignored in the model selection literature. In particular, if the global model fits the data does the selected model also fit? This appears to be a virtually unexplored question; we have not seen it rigorously addressed in the statistical literature. Post model-selection fit is an issue deserving of attention; we present here some ideas and results on the issue. Full-blown simulation evaluation would require a specific context of a data type and a class of models, data generation, model fitting, selection, and then application of an appropriate goodness-of-fit test (either absolute, or at least relative to the global model). This would be both time consuming and one might wonder if the inferences would generalize to other contexts.

A simple, informative shortcut can be employed to gain insights into the relative fit of the selected best model compared to a global model assumed to fit the data. The key to this shortcut is to deal with a single sequence of nested models, $g_1 \subset \cdots \subset g_i \subset g_{i+1} \subset \cdots \subset g_R$. It suffices that each model increments by one parameter, i.e., $K_{i+1} = K_i + 1$, and K_1 is arbitrary; $K_1 = 1$ is convenient as then $K_i = i$. In this context

$$\mathrm{AIC}_i = \mathrm{AIC}_{i+1} + \chi^2_1(\lambda_i) - 2$$

and

$$\mathrm{BIC}_i = \mathrm{BIC}_{i+1} + \chi^2_1(\lambda_i) - \log(n)\,,$$

where $\chi^2_1(\lambda_i)$ is a noncentral chi-square random variable on 1 degree of freedom with noncentrality parameter λ_i. In fact, $\chi^2_1(\lambda_i)$ is the likelihood ratio test statistic between models g_i and g_{i+1} (a type of relative, not absolute, goodness-of-fit test). Moreover, we can use $\lambda_i = n\lambda_{1i}$ where nominally λ_{1i} is for sample size 1. These λ are the parameter effect sizes and there is an analogy between them and the K-L distances here: the differences $I(f, g_i) - I(f, g_{i+1})$ are analogous to and behave like these λ_i.

Building on these ideas (cf., Burnham and Anderson 2002, pp. 412-414) we get

$$\mathrm{AIC}_i = \mathrm{AIC}_R + \sum_{j=i}^{R\text{-}1} (\chi_1^2(n\lambda_{1j}) - 2) ,$$

for for AIC_c

$$\mathrm{AIC}_{ci} = \mathrm{AIC}_{cR} + \sum_{j=i}^{R\text{-}1} \left[\chi_1^2(n\lambda_{1j}) - 2 + \frac{2K_i(K_i+1)}{n-K_i-1} - \frac{2(K_i+1)(K_i+2)}{n-K_i-2} \right] ,$$

and

$$\mathrm{BIC}_i = \mathrm{BIC}_R + \sum_{j=i}^{R\text{-}1} (\chi_1^2(n\lambda_{1j}) - \log(n)) .$$

To generate these sequences of model selection criteria in a coherent manner from the underlying “data" it suffices to, for example, generate the AIC_i based on the above and then use

$$\mathrm{AIC}_{ci} = \mathrm{AIC}_i + \frac{2K_i(K_i+1)}{n-K_i-1} ,$$

and

$$\mathrm{BIC}_i = \mathrm{AIC}_i - 2K_i + K_i \log(n)$$

to get the AIC_{ci} and BIC_i. Because only differences in AIC_c or BIC values matter it suffices to set AIC_R to a constant. Thus for specified R, K_1, n and λ_{1i} we generate the needed $R - 1$ independent noncentral chi-square random variables. Then we compute a realization of the sequences of AIC and BIC values for the underlying nested model sequence. We can then determine the best model under each model selection criterion.

If model g_h is selected as best under a criterion, for $h < R$, then the usual goodness-of-fit test statistic (for fit relative to the global model g_R) is

$$\chi_v^2 = \sum_{j=h}^{R\text{-}1} \chi_1^2(n\lambda_{1j}) ,$$

with degrees of freedom $v = K_R - K_h$ ($= R - h$ when $K_i = i$). Hence, we can simulate having one set of data, doing both AIC, or AIC_c, and BIC model selection for that data and then check the goodness-of-fit of each selected best model, relative to the baseline global model g_R. The results apply to discrete or continuous data, but do assume “large" *n*.

These simulations generate a lot tabular information, so we present below only a typical example. In general, we recommend the interested reader run their own simulations (they are easy to do and run quickly; SAS© code for doing this is available from KPB). We have done a lot of such simulation to explore primarily one question: after model selection with AIC or BIC does the selected model always fit, as judged by the usual likelihood ratio statistic *P*-value that tests g_R versus the selected model (this test ignores that a selection process was done)? Also, do the results differ for AIC versus BIC? We found that for large enough *n*, so that AIC_c and AIC are nearly the same, then for a Figure 1 scenario (i.e. realistic data), (1) the AIC selected model always fits relative to the global model, and (2), the BIC selected model too often (relative to the α-level of the test) fails to fit the data. Under a scenario such as in Figure 2 the BIC selected model generally fits the data; GOF results for AIC model selection are about the same for all three scenarios.

To be more precise let $\alpha = 0.05$ so we say the selected model fits if the (relative) goodness-of-fit test *P*-value > 0.05. Then for the AIC selected model we almost always find $P > 0.05$. However, for the BIC selected model, under tapering effects, the probability that $P < 0.05$ occurs can be much higher than the nominal $\alpha = 0.05$. For example, let $R = 10$, $K_i = i$, and $\lambda_1(1)$ to $\lambda_1(10)$ be 0.3, 0.2, 0.15, 0.1, 0.05, 0.025, 0.01, 0.005, 0.001, 0.0003, respectively (mimics Figure 1). Table 3 gives some of these goodness-of-fit results for AIC_c and BIC under this scenario for a few values of *n*.

table 3 about here

In Table 3 the key column is column three. It is the relative frequency at which the selected best model g_h did not fit relative to model 10 (the global model here), in the sense that its GOF *P*-value was ≤ 0.05. In calculating this statistic if the selected model was model

10 we assumed the model fit. Hence, the lack of fit statistic in Table 10 column three would be larger if it were only for when the selected best model was model 1 through 9. Column four of Table 3 gives the frequency, out of 10,000 trials, wherein the best model was one of models 1 to 9. These GOF results are striking. The model selected as best by AIC_c (which is not really different here from AIC at $n \geq 200$) rarely leads to a GOF *P*-value $< \alpha = 0.05$ for $n \geq 100$. The best BIC model often fails to fit, relative to model 10, in terms of its GOF *P*-value being ≤ 0.05 (e.g, GOF failure rate of 0.217 at $n = 200$ here). Columns 5 to 9 of Table 3 provide further summaries of these GOF *P*-values when the selected best model was 1 through 9.

These results are not atypical under tapering effects. For the Figure 2 scenario that favors BIC, GOF for the BIC selected model comes much closer to nominal levels. Thus again, operating characteristics of AIC and BIC depend on the underlying scenario about reality versus the model set. What should we make of such results for the tapering effects case? Is it bad that the AIC-best model always fits: is it overfitting? Is it bad that the BIC-best model fails to fit at a much higher rate than the α-level: is it underfitting? We do not know because to have evidence about the matter we need to have a context and actual parameters estimated and look at mean square errors and confidence interval coverage (see Burnham and Anderson 2002, pp. 207-223).

We make four comments on the issues. First, as regards a perception of “overfit" by AIC, surely when one deliberately seeks a good model for analysis of data one is seeking a good fit. Thus if the global model fits, we think one would expect the best model, under a selection criterion, to also fit. Heuristically, it is a strange model selection criterion that often selects a best model that fits poorly; AIC does not do this. However, we also claim the best model often allows some bias in estimates, which could be analogous to some lack of fit. Therefore, second, with regard to BIC the degree of lack of fit may not matter – we do not know so do not claim it matters. Third, model averaged inference further renders the GOF issue somewhat moot because all the models are being considered, not just the best model.

Forth, these observations and issues about fit reinforce to us that model selection procedures should be judged on their inferential operating characteristics, such as predictive mean square error and interval coverage under realistic scenarios for generation of data.

7. DISCUSSION AND CONCLUSIONS

The context of classical model selection proceeds in 4 steps:

(1) the goal is model-based inference from data, and

(2) there is a set of R relevant models, but no certainty about which model should be used; hence,

(3) a data-based choice is made among these (perceived as) competing models, and

(4) then inference is made from this one selected model as if it were a priori the only model fit to the data.

Steps 1 and 2 are almost universal in model-based inference. Step 3 begins a flawed inference scenario, in particular the implicit assumption that inference must be based on a single model is not justified by any philosophy or mathematics.

To avoid the pitfalls inherent in step 4 we must conceptualize model selection to mean, and be, multimodel inference. The new step 3 should be

(3) there is a data-based assignment of model weights that sum to 1.0; the weight for model g_i reflects the evidence or information concerning model g_i (uncertainty of model g_i in the set of R models).

The old step 3 is subsumed in this new step 3 because the model with the highest weight is the model that would be selected as the single best model. But now we avoid many of the problems that stem from old step 4 by using a new step 4.

(4) Based on the model weights, and the results and information from the R fitted, models we use multimodel inference, in some or all of its myriad forms and methods.

Model selection should be viewed as the way to obtain model weights, not just a way to select only one model (and then ignore that selection occurred).

Among other benefits of this approach it effectively rules out null hypothesis testing as a basis for model selection, because multimodel inference forces a deeper approach to model selection. It means we must have an optimality criterion and selection (weight assignment) theory underlying the approach. Potential users should not reject or ignore multimodel inference just because it is relatively new, especially when based on AIC. There is a sound philosophical basis and likelihood framework for AIC, based on on Kullback-Leibler information theory, which itself has a deep foundation.

An important issue about model selection based on K-L information is that AIC as such is a large sample approximation (relative to the maximum K for the model set) to the needed criterion. A second order bias adjustment is needed when n/K is too small, say ≤ 40. While AIC_c is not unique as providing the needed small-sample version of AIC, we recommend it for general use, and indeed the evidence is that it performs well. Much confusion and misinformation has resulted in the model selection literature when investigators have done simulation evaluations using AIC when they should have used AIC_c (Anderson and Burnham 2002, pp. 287-293).

A compatible, alternative view of AIC is that it arises from a Bayesian derivation based on the BIC statistic and a savvy prior probability distribution on the R models. That prior depends on both n and K_i ($i = 1, \ldots, R$) in a manner consistent with the information-theoretic viewpoint that the data at hand surely reflect a range of tapering effects based on a complex reality, rather than arising from a simple true model, with no tapering effects, that is in the model set.

The model selection literature often errors by considering that AIC and BIC selection are directly comparable, as if they had the same objective target model. Their target models are different (Reschenhofer 1996). The target model of AIC is one that is specific for the sample size at hand: it is the fitted model that minimizes expected estimated Kullback-Leibler information loss when fitted model g_r is used to approximate full reality, f. This target model

changes which sample size. Moreover, in this overall philosophy, even the set of models is expected to be changed if there are large changes in *n*.

The classical derivation of BIC assumed there was a true model, independent of *n*, that generated the data, it was a model in the model set, and this true model was the target model for selection by BIC. However, selection of this true model with probability 1 only occurs in the limit as *n* gets very large and in taking that limit the model set is kept fixed. The original derivation of BIC has been relaxed, wherein we realize that such convergence only justifies an inference of a quasi-true model (the most parsimonious model closest in K-L information to truth, f). Even within the Bayesian framework not all practitioners subscribe to BIC for model selection (some Bayesians do not believe in model selection at all). In particular, we note the recent development of the deviance information criterion (DIC) by Spiegelhalter et al. (2002). As these authors note, DIC behaves like AIC, not like BIC which is one reason they prefer DIC (it avoids the defects of BIC model selection).

Given that AIC can be derived from the BIC approximation to the Bayes factor the distinction between AIC vs. BIC model selection becomes one about the prior on models: $q_i = 1/R$ for BIC, or for AIC the K-L prior of section 4 (formulae 3a, 3b). This latter prior is a savvy prior, by which we mean that the expected number of parameters that can be estimated with useful precision depends on *n* and *K* (which are known a priori). Thus for a savvy prior, in general, q_i becomes a function of *n* and K_i, say $q_i(K_i, n)$, and we think in terms of prior $E(K) = n/m$, for some *m*, perhaps in the 10 or 15 range. Fitting a model with too few parameters wastes information. With too many parameters in a model some or all (with typical correlated observational data) of the estimated parameters are too imprecise to be inferentially useful.

Objective Bayesian analysis with a single model uses an uninformative (vague) prior such as U(0, 1) on a parameter θ if $0 < \theta < 1$. This turns out to be quite safe, sort of "innocent" one might say (no lurking unexpected consequences). So presumably it seemed natural, objective, and innocent when extending Bayesian methods to model selection to

assume a uniform prior on models. However, we now know that this assumption has unexpected consequences (it is not innocent), as regards to the properties of the resultant model selection procedure. Conversely, there is a rationale for considering that the prior on models ought to depend on n and K, and so doing produces some quite different properties of the selection method as compared to use of $1/R$. The choice of the prior on models can be important in model selection and we maintain q_i should usually be a function of n and K.

Whereas the best model selected by either BIC or AIC can be distinctly different and hence suggest partially conflicting inferences, model averaged inference diminishes the perceived conflicts between AIC and BIC. In general, we have seen robustness of inference to variations in the model weights for rational choices of these weights. For this reason we think there is little need to seek alternative savvy priors to the K-L prior.

Several lines of thinking motivate us to say the comparison of AIC and BIC model selection ought to be based on their performance properties such as mean square error for parameter estimation (includes prediction) and confidence interval coverage: tapering effects or not, goodness-of-fit issues, derivation of theory is irrelevant as it can be frequentist or Bayes. When any such comparisons are done the context must be spelled out explicitly because results (i.e., which method “wins") depend on context (e.g, Figures 1-3). Simulation evaluations should generate realistically complex data, should use AIC_c, and should use multimodel inference, hence go well beyond the traditional single best model approach.

We believe that data analysis should routinely be considered in the context of multimodel inference. Formal inference from more than one (estimated best) model arises naturally from both a science context (multiple working hypotheses) and a statistical context (robust inference, while making minimal assumptions). The information-theoretic procedures allowing multimodel inference are simple, both in terms of understanding and computation, and, when used properly, provide inferences with good properties, e.g., as regards predictive mean squared error and achieved confidence interval coverage. Multimodel inference goes beyond the concepts and methods noted here; we give a richer account in Burnham and

Anderson (2002). Model selection bias and model selection uncertainty are important issues that deserve further understanding. Multimodel inference is an new field where additional, innovative research and understanding is needed and we expect a variety of important advances to appear in the years ahead.

REFERENCES

Akaike, Hirotugu. 1973. “Information Theory as an Extension of the Maximum Likelihood Principle." Pp. 267-281, in Second International Symposium on Information Theory, edited by B. N. Petrov and F. Csaki. Budapest: Akademiai Kiado.

Akaike, Hirotugu. 1974. “A New Look at the Statistical Model Identification." IEEE Transactions on Automatic Control AC-19:716-723.

Akaike, Hirotugu. 1981. “Likelihood of a Model and Information Criteria." Journal of Econometrics 16:3-14.

Akaike, Hirotugu. 1983. “Information Measures and Model Selection." International Statistical Institute 44:277-291.

Akaike, Hirotugu. 1985. “Prediction and Entropy." Pp. 1-24, in A Celebration of Statistics, edited by Anthony C. Atkinson and Stephen E. Fienberg. New York: Springer-Verlag.

Akaike, Hirotugu. 1992. “Information Theory and an Extension of the Maximum Likelihood Principle." Pp. 610-624, in Breakthroughs in Statistics, vol. 1, edited by Samuel Kotz and Norman L. Johnson. London: Springer-Verlag.

Akaike, Hirotugu. 1994. “Implications of the Informational Point of View on the Development of Statistical Science." Pp. 27-38, in Engineering and Scientific Applications, vol. 3. Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach, edited by Hamparsum Bozdogan. Dordrecht, Netherlands: Kluwer Academic Publishers.

Anderson, David R. and Kenneth P. Burnham 2002. “Avoiding Pitfalls When Using Information-Theoretic Methods." Journal or Wildlife Management 66:910-916.

Azzalini, Adelchi. 1996. Statistical Inference Based on the Likelihood. London: Chapman and Hall.

Boltzmann, Ludwig. 1877. “Uber die Beziehung Zwischen dem Hauptsatze der Mechanischen Warmetheorie und der Wahrscheinlicjkeitsrechnung Respective den Satzen uber das Warmegleichgewicht." Wiener Berichte 76:373-435.

Breiman, Leo. 1992. “The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: *X*-fixed Prediction Error." Journal of the American Statistical Association 87:738-754.

Breiman, Leo. 2001. “Statistical Modeling: the Two Cultures." Statistical Science 26:199-231.

Buckland, Steven. T., Kenneth P. Burnham, and Nicole H. Augustin. 1997. “Model Selection: an Integral Part of Inference." Biometrics 53:603-618.

Burnham, Kenneth P. and David R. Anderson. 1998. Model Selection and Inference: A Practical Information-Theoretical Approach. New York: Springer-Verlag.

Burnham, Kenneth P. and David R. Anderson. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach. 2d ed. New York: Springer-Verlag.

Cavanaugh, Joseph E. and Andrew A. Neath. 1999. “Generalizing the Derivation of the Schwarz Information Criterion." Communication in Statistics Theory and Methods 28:49-66.

Chamberlin, Thomas. 1890. “The Method of Multiple Working Hypotheses." Science 15:93-98.

Chamberlin, Thomas. 1965. “The Method of Multiple Working Hypotheses." Science 148:754-759. (reprint of the 1890 paper in Science).

deLeeuw, Jan. 1992. “Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle." Pp. 599-609, in Breakthroughs in Statistics, vol. 1, edited by Samuel Kotz and Norman L. Johnson. London: Springer-Verlag.

Edwards, Anthony W. F. 1992. Likelihood, expanded ed. Baltimore, MD: Johns Hopkins University Press.

Forster, Malcolm. R. 2000. “Key Concepts in Model Selection: Performance and Generalizability." Journal of Mathematical Psychology 44:205-231.

Forster, Malcolm R. 2001. “The New Science of Simplicity." Pp. 83-119, in Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple, edited by Arnold Zellner, Hugo A. Keuzenkamp, and Michael McAleer. Cambridge, UK: Cambridge University Press.

Forster, Malcolm R. and Elliott Sober. 1994. “How to Tell Simpler, More Unified, or Less *ad hoc* Theories will Provide more Accurate Predictions." British Journal of the Philosophy of Science 45:1-35.

Gelfand, Alan and Dipak K. Dey. 1994. “Bayesian Model Choice: Asymptotics and Exact Calculations." Journal of the Royal Statistical Society, Series B 56:501-514.

Gelman, Andrew, John C. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. Bayesian data analysis. New York: Chapman and Hall.

Hand, David J., and Veronica Vinciotti. 2003. “Local Versus Global Models for Classification Problems: Fitting Models where it Matters." The American Statistician 57:124-131.

Hansen, Mark H. and Charles Kooperberg. 2002. “Spline Adaptation in Extended Linear Models." Statistical Science 17:2-51.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky. 1999. “Bayesian Model Averaging: a Tutorial (with Discussion)." Statistical Science 14:382-417.

Hurvich, Clifford M. and Chih-Ling Tsai. 1989. “Regression and Time Series Model Selection in Small Samples." Biometrika 76:297-307.

Hurvich, Clifford M. and Chih-Ling Tsai. 1995. “Model Selection for Extended Quasi-Likelihood Models in Small Samples." Biometrics 51:1077-1084.

Johnson, Roger W. 1996. “Fitting Percentage of Body Fat to Simple Body Measurements." Journal of Statistics Education v.4,n.1 (an e-journal), http://www.amstat.org/publications/jse/v4n1/datasets.johnson.html

Kass, Robert E. and Adrian E. Raftery. 1995. “Bayes Factors." Journal of the American Statistical Association 90:773-795.

Key, Jane T., Luis R. Pericchi, and Adrian F. M. Smith. 1999. “Bayesian Model Choice: What and Why?" Pp. 343-370, in Bayesian Statistics 6, edited by José M. Bernardo, James O. Berger, A. Philip Dawid, and Adrian F. M. Smith. Oxford UK: Oxford University Press.

Kullback, Soloman and Richard A. Leibler. 1951. “On Information and Sufficiency." Annals of Mathematical Statistics 22:79-86.

Lahiri, Partha (editor). 2001. Model Selection. Beachwood, Ohio: Lecture Notes – Monograph Series, Institute of Mathematical Statistics.

Lehman, Eric L. 1990. “Model Specification: the Views of Fisher and Neyman, and Later Observations." Statistical Science 5:160-168

Linhart, H. and Walter Zucchini. 1986. Model Selection. New York. Wiley.

Lukacs, P. L., Kenneth P. Burnham and David R. Anderson. (in review). “Model Averaging in Linear Regression: Revisiting the Problem of Spurious Effects." Technometrics.

McQuarrie, Alan D. R. and Chih-Ling Tsai. 1998. Regression and Time Series Model Selection. Singapore: World Scientific Publishing Company.

Meyer, Mary C. and Purushottam W. Laud. 2002. “Predictive Variable Selection in Generalized Linear Models." Journal of the American Statistical Association 97:859-871.

Parzen, Emmanuel, Kunio Tanabe, and Genshiro Kitagawa. (Eds.). 1998. Selected Papers of Hirotugu Akaike. New York: Springer-Verlag.

Raftery, Adrian E. 1995. “Bayesian Model Selection in Social Research (with Discussion)." Sociological Methodology 25:111-195.

Raftery, Adrian E. 1996. “Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Regression Models." Biometrika 83:251-266.

Reschenhofer, Erhard. 1996. “Prediction With Vague Prior Knowledge." Communications in Statistics – Theory and Methods 25:601-608.

Royall, Richard M. 1997. Statistical Evidence: A Likelihood Paradigm. London: Chapman and Hall.

Stone, Mervyn. 1974. “Cross-Validatory Choice and Assessment of Statistical Predictions (with Discussion)." Journal of the Royal Statistical Society, Series B 39:111-147.

Stone, Mervyn. 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.

Schwarz, Gideon. 1978. “Estimating the Dimension of a Model." Annals of Statistics 6: 461-464.

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelita van der Linde. 2002. “Bayesian Measures of Model Complexity and Fit." Journal of the Royal Statistical Society, Series B 64:1-34.

Sugiura, Nariaki 1978. “Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections." Communications in Statistics, Theory and Methods A7:13-26.

Takeuchi, Kei. 1976. “Distribution of Informational Statistics and a Criterion of Model Fitting." Suri-Kagaku (Mathematic Sciences) 153:12-18. (In Japanese).

Wasserman, Larry. 2000. “Bayesian Model Selection and Model Averaging." Journal of Mathematical Psychology 44:92-107.

Weakliem, David L. 1999. “A Critique of the Bayesian Information Criterion for Model Selection." Sociological Methods & Research 27:359-397.

Williams, David. 2001. Weighing the Odds: A Course in Probability and Statistics. Cambridge, UK: Cambridge University Press.

Table 1. Effects, as $\beta/\mathrm{se}(\hat{\beta})$, in the models used for Monte Carlo simulation based on the body fat data to get predictive mean square error results by model selection method (AIC_c or BIC) and prediction strategy (best model or model averaged); model *i* has the effects listed on lines 1 to *i* and it's remaining β are 0.

i	$\beta/\mathrm{se}(\hat{\beta})$	Variable *j*
1	11.245	6
2	- 3.408	13
3	2.307	12
4	- 2.052	4
5	1.787	8
6	- 1.731	2
7	1.691	1
8	- 1.487	7
9	1.422	11
10	1.277	10
11	- 0.510	5
12	- 0.454	3
13	0.048	9

Table 2. Ratios of predictive mean square error (PMSE) based on Monte Carlo simulation patterned after the body fat data, 10,000 independent trials for each generating model; margin of error for each ratio is 3%; generating model *i* has exactly *i* effects, ordered largest to smallest for models 1 to 13 (see Table 1 and text for details).

	PMSE ratios of $AIC_c \div BIC$		PMSE Ratios for model averaged ÷ best	
Generating model *i*	Best model	Model averaged	AIC_c	BIC
1	2.53	1.97	0.73	0.94
2	1.83	1.51	0.80	0.97
3	1.18	1.15	0.83	0.85
4	1.01	1.05	0.84	0.81
5	0.87	0.95	0.84	0.77
6	0.78	0.88	0.87	0.77
7	0.77	0.86	0.86	0.77
8	0.80	0.87	0.85	0.78
9	0.80	0.87	0.85	0.78
10	0.72	0.81	0.85	0.75
11	0.74	0.82	0.84	0.76
12	0.74	0.81	0.84	0.76
13	0.74	0.82	0.83	0.75

Table 3. Simulation of goodness-of-fit (GOF) results after model selection (see text for details) for $R = 10$ nested models, $K_i = i$, effects $\lambda_1(1)$ to $\lambda_1(10)$ as 0.3, 0.2, 0.15, 0.1, 0.05, 0.025, 0.01, 0.005, 0.001, 0.0003, respectively; 10,000 trials at each n, $\alpha = 0.05$; model g_{10} was consider to always fit so results on GOF relate only to models g_i , $i < 10$.

Sample size n	Selection method	Rel. freq. not fitting	Freq. of $i \leq 10$	Mean of GOF P	Percentiles of P-value 1	5	10	25
50	AIC_c	0.026	9961	0.470	0.030	0.073	0.118	0.246
	BIC	0.115	9995	0.352	0.006	0.022	0.044	0.117
100	AIC_c	0.004	9809	0.511	0.063	0.120	0.171	0.296
	BIC	0.159	9995	0.470	0.003	0.014	0.030	0.087
200	AIC_c	0.004	9569	0.531	0.096	0.155	0.202	0.328
	BIC	0.217	9997	0.273	0.002	0.009	0.019	0.062
500	AIC_c	0.000	9178	0.546	0.127	0.178	0.224	0.345
	BIC	0.281	9992	0.236	0.001	0.005	0.011	0.041
1000	AIC_c	0.000	8662	0.537	0.136	0.176	0.218	0.339
	BIC	0.320	9978	0.227	0.001	0.004	0.009	0.035
10000	AIC_c	0.000	3761	0.448	0.159	0.171	0.187	0.244
	BIC	0.509	9295	0.135	0.000	0.001	0.002	0.009

Figure 1. Values of Kullback-Leibler information loss, $I(f, g_r(\cdot \mid \theta_o)) \equiv nI_1(f, g_r(\cdot \mid \theta_o))$, illustrated under tapering effects for 12 models ordered by decreasing K-L information; sample size n, hence the y-axis, is left unspecified; this scenario favors AIC-based model selection.

Figure 2. Values of Kullback-Leibler information loss, $I(f, g_r(\cdot \mid \theta_o)) \equiv nI_1(f, g_r(\cdot \mid \theta_o))$, illustrated when models 1 (simplest) to 12 (most general) are nested with only a few big effects; model 4 is a quasi-true model, models 5 to 12 are too general; sample size n, hence the y-axis, is left unspecified; this scenario favors BIC-based model selection.

Figure 3. Values of Kullback-Leibler information loss, $I(f, g_r(\cdot \mid \theta_o)) \equiv nI_1(f, g_r(\cdot \mid \theta_o))$, illustrated when models 1 (simplest) to 12 (most general) are nested with a few big effects (in model 4), then much smaller tapering effects (models 5 to 12); whether BIC or AIC is favored depends on sample size.

Drs. Kenneth P. Burnham and David R. Anderson work at the Colorado Cooperative Fish and Wildlife Research Unit at Colorado State University in Fort Collins. They are employed by the U.S. Geological Survey, Division of Biological Resources; they have graduate faculty appointments in the Department of Fishery and Wildlife Biology and teach a variety of quantitative graduate courses. They have worked closely together since 1973 where they met and worked together at the Patuxent Wildlife Research Center in Maryland. Much of their joint work has been in the general areas of capture-recapture and distance sampling theory. Their interest in model selection arose during research on the open population capture-recapture models in the early 1990s. They have jointly published 10 books and research monographs and 71 journal papers on a variety of scientific issues. Most relevant here is their 2002 Springer-Verlag book, Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.

Ken Burnham is a statistician and has over 32 years (post PhD) experience developing and applying statistical theory in several areas of the life sciences, especially ecology, fisheries, and wildlife. He is a Fellow of the American Statistical Association (1990).
David Anderson is a theoretical ecologist and has over 36 years working at the interface between biology and statistics. He received the Meritorious Service Award from the U.S. Department of the Interior and was awarded a senior scientist position in 1999.

Title Page

MULTIMODEL INFERENCE:

Understanding AIC and BIC in Model Selection

Kenneth P. Burnham

David R. Anderson

Colorado Cooperative Fish and Wildlife Research Unit (USGS-BRD)

Key words: AIC, BIC, model averaging, model selection, multimodel inference.

Sociological Methods and Research