Model weights and the foundations of multimodel inference: Comment

Kenneth P. Burnham[1,2] and David R. Anderson[3]

Key words: AIC, AICc, Bayesian inference, BIC, Model averaging, Model selection.

INTRODUCTION

Recently, Link and Barker (2006) (denoted LB here) promoted a Bayesian framework for multimodel inference, covering both model selection and model averaging. They provide a concise review of the Bayesian approach. They suggest the use of BIC but caution its use to compute approximate posterior model weights and they discuss and illustrate technical difficulties associated with Bayes factors. We find ourselves in agreement on several points and view LB as a contribution. We feel that several points deserve further comment and there are several important issues that need clarification such that readers will not be misinformed about our views on the information-theoretic approach.

---

[1] USGS, Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins, Colorado 80523 USA

[2] Corresponding author. Email: kenb@warnercnr.colostate.edu

[3] Applied Information Company, 2214 Merlot Court, Fort Collins, Colorado, 80528 USA

TRUTH VS. TRUE MODELS

LB's discussion of truth (full reality) and so-called true models obscures a long-term contention in the literature. We believe that few scientists would question the concept of "truth" -- this is *not* the issue. The focus of first concern is if a mathematical model can fully represent "truth" in its entirety; i.e., do "true models" exist in the biological sciences? The second and conditional issue asks if this true model is in the model set but the investigator does not know which one it is -- thus the need for some type of model selection to identify which model is the true one.

LB: p628 state "Most would concede that it is unlikely that Truth is in our model set." Of course not, only models are in the model set -- truth is not a model. One cannot have a model set such as $\{g_1, g_2, g_3, Truth, g_4, \ldots, g_R\}$. Truth cannot be in a model set – this statement obstructs clear thinking about this science philosophy issue. Rather, consider the model set $\{g_1, g_2, g_3, g_4, \ldots, g_R\}$; now ask if any one of these *models* is an exact representation of truth or full reality. That is, "is the (supposed) true *model* in the model set?" It is critical here to clearly distinguish between the concepts of truth or full reality and a mathematical model.

We have argued (Burnham and Anderson 2002: 20) that there are no true models in the biological sciences, certainly not in the sense that real data literally are produced from such a model. Real data do not come from (parametric) mathematical models as is often assumed in some statistical literature. The unfounded notion of a true model producing the data is the thinking of a mathematician not of an applied scientist. Statistical theorists often perform Monte Carlo simulation studies where "data" are generated from a mathematical model using pseudorandom numbers. In this computer sense, there is a "true computer model" that generated the "data" – we refer to this as a generating model; it is assumed to be given (i.e., the specific model form and its parameters are chosen for study and known before the "data" are generated). People then state formally that they assume (real) data come from a parametric model! This is nonsense. Empirical data come from measurements or counts from full reality, hence from some ecological system of interest. If a proper experimental design or sampling protocol is used, we might hope to obtain data that partially represents truth about the system under study and can therefore lead to a partial understanding of the system (i.e., reality). We must not believe that actual data come from mathematical models or that a model can ever be a perfect representation of truth or full reality. Models are approximations by definition (see Burnham and Anderson 2002:20-26 and 289-293.

LB: p2626 and 2628 claim that standard statistical models are an *exact* depiction of the process that generated the data. They seem to reason that this exactness pervades much of statistics and, therefore, model selection can proceed as if this was a standard. We view this very differently: take a simple quadratic regression model as an illustration,

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2 \ .$$

Under certain standard assumptions, there are well known estimators of the model parameters and these are optimal in some mathematically rigorous sense. While countless applications have used this model; we doubt if anyone thought that this was an exactly true model of some system or process. Instead, people believed that this model might be useful in the analysis of real data; that is, it might serve as a handy approximation of the information contained in the data. Surely no one thought that their real data came from this model.

The matter of exactly true models is important and goes well beyond semantics (Bozdogan 1987). After many words and personal opinion, but with little rigor, logic, or justification LB: p2629 conclude that "Bayesian multimodel inference uses "truth in the model set" as a model itself, rather than a statement of reality." We think there are deep issues here that need to be more clearly addressed.

There are still further issues with respect to the usual Bayesian prior on models – p_i is the probability that model *i* is exactly true. We stress exactly because if a model is "nearly true" then we are back to acknowledging that models are approximations. Given this definition of a Bayesian prior on models then it is clear that the posterior distribution must also relate to truth..

AICc SHOULD HAVE BEEN THE BASIS FOR COMPARISON

LB's paper makes several comments concerning AIC and this is unfortunate in that AICc should have been used, not AIC. Akaike (1973) derived his AIC as an expectation of estimated Kullback-Leibler information using an asymptotic bias correction of the maximized log-likelihood. Suguira (1978) and several others since have provided a second order bias correction, often termed AICc. The fact that AIC tends to select models that are over-fit (i.e., too many parameters) for small sample sizes has been known for decades, thus LB's comments were predicted. AICc has been available for 28 years and should be well known to statisticians studying AIC-based model selection. LB discuss only AIC and this leads them to noting the tendency of AIC weights to favor complex models; however, AICc does not share this tendency (but this does not affect their example as that sample size is quite large (R. Barker, pers. comm.). Their statement might lead some readers to mistakenly think that BIC is somehow always "better" for the analysis of real data.

ACCOUNTING FOR PARAMETER ESTIMATION

LB: p2629 are mistaken when they state, "…there is no formal accounting for the uncertainty in parameter estimation in the definition of AIC weights." Burnham and Anderson (2002, Section 7.2) provide Akaike's derivation involving a second expectation over the estimated parameters. This material makes it clear that AIC and AICc are based on an expectation of Kullback-Leibler information. We have not seen this mistake elsewhere in the literature and cannot find the basis for their claim. We show (Burnham and Anderson, Chapter 7) that the maximum likelihood estimator of θ is consistent for the Kullback-Information minimizer θ_o. While a rather technical issue, this is important.

BAYESIAN METHODS ARE HARDLY ALONE!

LB: p2629 assert, "…model weighting has no compelling epistemological foundation outside of the Bayesian paradigm." We take strong exception to this statement (and a similar statement on page 2631) because Kullback-Leibler information theory provides a fundamental theory for general statistical inference. Akaike (1973) suggested his information-theoretic approach be considered an extension of likelihood theory. Methods developed over the past 30 years based on K-L information have seen little controversy; this is in contrast to the 250 year old Bayesian approach that has seen wide opposition to its subjectivity and other issues centered on both priors-on-parameters and, recently, priors-on-models. Beyond that, we must ask for the basis for such a sweeping statement. We have studied the model selection literature carefully since 1990 and have not seen such an unwarranted assertion.

Much of our work has found productive interplay between information-theoretic and Bayesian approaches to inference (e.g., Burnham and Anderson 2004). Often there is agreement in broad concepts or performance measures. BIC can be derived from frequentist principles while AIC can be derived from Bayesian principles. The fact that the information-theoretic approaches represent a simple alternative does not diminish the value of the complex and computer intensive Bayesian methods. It is common that "good" statistical methods can be considered either Bayesian or frequentist (e.g., random effects models).

We argue in Burnham and Anderson (2004) that these issues are generally not Bayesian vs. non-Bayesian, rather the issue deals with the notion of exactly true models and the meaning of priors on these models. Perhaps it would be helpful to have a real example where one had *R* models, one of which was a model perfectly representing full reality, but the investigator did not know which of the models was that true model. The tone of LB's paper, to us, seems to promote (only) the Bayesian approach (including Bayes factors) and we view this as counter productive. There are deeper science philosophy issues here than any particular approach being "Bayesian." These include the goal of inference: is it to find the true (or quasi-true) model or is it prediction, including concepts of sample size and bias vs. variance trade-offs leading to parsimony.

It is important to realize that it is the *fitted* model that is the cornerstone of statistical inference. Bayesian approaches attempt to identify the true (or quasi-true, see Burnham and Anderson 2004) model, $g(\theta_o)$ including its parameters, while approaches based on Kullback-Leibler information select a parsimonious predictive model $g(\hat{\theta}_o)$ where the parameters are *estimated*.

THE ISSUE OF APPROXIMATIONS AND ASYMPTOTICS

LB: p2634 state "The beauty of the Bayesian calculus is in its transparency and precision: posterior distributions are exactly determined by the specification of models for data, priors for parameters, and prior model weights; there is no need for approximations of unknown precision, no need for dubious asymptotics, no need for buried assumptions." In reality, the posterior distributions are approximations, subject to asymptotic convergence issues stemming from very intricate MCMC computations and can hardly be called exact. Their statement is incorrect and tends to mislead the reader. For example, Givens and Hoeting (2005:200-201) note that " In practice, however, it is necessary to determine when the chain has run significantly long so that it is reasonable to believe that the output adequately represents the target distribution and can be used reliably for estimation. Unfortunately, MCMC methods can sometimes be quite slow to converge, requiring extremely long runs, especially if the dimensionality of **X** is large. Further, it is often easy to be mislead when using the MCMC algorithm output to judge whether convergence has approximately been obtained." LB used a small data set on brown trout (*Salmo trutta*); this simple example had only 5 models, the largest model had only 6 parameters but the analysis used 5,000,000 MCMC trials and took 6.5 hours of computer time. "Exactness" seems strained for this almost trivial example. Givens and Hoeting (2005:201-212) provide several useful approaches for assuring that the approximations of the target distribution are adequate; they note "The art of MCMC lies in the construction of a suitable chain."

BIC is clearly derived from asymptotics and approaches its target (the true model that is assumed to be in the model set) from below, often leading to the selection of under-fitting models when sample sizes are less than very large. Again, the dependence on the true model being in the model set must surely be viewed as a "buried assumption."

Surely LB would recognize the dubious asymptotics in using AIC, so we question why LB used this criterion. AICc allows a second order bias correction and is useful in model selection when sample sizes are small, moderate, or large. Analysts should avoid asymptotic methods if simple alternatives exist.

BIC SELECTION OUTCOMES SEEM INCONGRUOUS

LB do not address the issue that BIC not uncommonly places high posterior model weights on models that do not fit (Burnham and Anderson 2002:309-319, Burnham and Anderson 2004: 293-297). We have seen this anomaly in both applied problems and in computer simulation results. This undesired outcome may reflect the inadequacy of the underlying theory when the true model is assumed to be in the model set and the sample size is assumed to be asymptotic but, in fact, the true model is not in the set and the sample size may be fairly small. More work is needed to better understand this issue.

Theory often makes use of asymptotic assumptions, but in application the asymptotics are realized for samples that are quite small. For example, many parameter estimation issues can be quite satisfactory although sample size might be only 30-70 (unless parameter estimates are on a boundary). This level of asymptotics is often not the case in using BIC in model selection; we have found the need for very large samples for BIC to show consistency for its target model (Burnham and Anderson 2002:292).

Other incongruities are illustrated by the capture-recapture data on European dipper (*Cinclus cinclus*) (Lebreton et al. 1992). The research issue concerns the effect, if any, of a flood on annual apparent survival probabilities of this streamside bird. Consider two models $\{\varphi, p\}$ and $\{\varphi_n, \varphi_f, p\}$ where φ is survival probability and p is recapture probability. The first model has only 2 parameters while the second has 3 parameters because the survival years of the study are partitioned into flood (parts of 2 years) and non-flood (the remaining years). The first model is nested within the second model with a LRT statistic of 6.735; assuming this is chi-squared on 1 degree of freedom, we obtain a P-value of 0.0095. Many people would take this as strong evidence of a flood effect on annual apparent survival. AICc (and DIC) also select the flood model with 3 parameters as the better of the two. The two model probabilities were the same under AICc and DIC: 0.087 vs. 0.913 for model 1 and 2, respectively. However, BIC gives the model probabilities as 0.37 and 0.63, respectively and this does not seem consistent with all the other evidence. Moreover, we ran a fully Bayesian analysis (MCMC, vague priors, 4,000 tuning trials, 10,000 burn-in trials, and 100,000 trials to sample the posterior distribution) to compute the posterior probability under model 2 that the difference and $\varphi_{n} - \varphi_{f} > 0$. This probability was 0.9956, again in reasonable agreement with the other approaches except BIC (note this posterior model probability is approximate to only 2-3 significant places even though we used 100,000 MCMC trials). Clearly, the meaning of the posterior model probabilities differs between AICc and DIC (best predictive, including estimation) vs. BIC (true or quasi-true model). Still, such incongruous results seem bothersome, particularly in such a simple case as this.

SUBJECTIVE AND ARBITRARY PRIOR PROBABILITIES ON MODELS

In relation to model priors, LB: p2631 state “It is, to our mind, far better to lay subjective choices out on the table and to present a mathematically precise analysis, than to ignore automatic choices in approximate analyses, and to mistake arbitrariness for objectivity.” While we respect their view, we question the difference between “subjective” and “arbitrariness” when faced with assigning prior probabilities on models. If one decides, for example, to subjectively place 0.428 probability on model 3, how is that not arbitrary? What theory guides a rational decision to make this prior probability 0.428? Our view is that if one starts with a fundamental theory and target (i.e., K-L information or entropy), then there is little need for either subjectivity or arbitrary decisions that arise in either BIC or some other forms of Bayesian inference. Good science is partly about minimizing subjectivity (instead of laying out several subjective choices and arbitrarily picking one).

If one were to "lay subjective choices out on the table" then how might an objective analysis be done when controversies are involved? Science must strive for more than a theory-less, arbitrary choice among subjective alternatives. In conflicts or controversies each party would likely have their own brand of subjectivity and arbitrariness. It is not clear to us what LB mean by a "mathematically precise analysis" as precision is compromised in cases where the (subjective) prior is not swamped by the data. In this case, different results stem from different priors on either parameters or models.

Going further (e.g., Jeffreys 1973), are not Bayesian model priors supposed to come from prior "beliefs"? Surely is it logical to know something of the sample size (n) and the 'size' of the model (K) when (subjectively) assigning prior probabilities on models. This thinking leads to savvy model priors in a Bayesian context (Burnham and Anderson 2004:280-283) and helps our understanding as to *why* AICc performs relatively well across a wide spectrum of applications. Asking "what priors would have to be assumed to make AIC fall out of a Bayesian framework?" gives us a new basis of justifying Bayesian model priors and lessening their subjectivity. Savvy priors invoke common sense about the role of sample size and model dimension.

Consider a consulting situation where two investigators seek advice in the analysis of real data on some system or process. Both investigators have three models, having 3, 22 and 76 parameters, respectively. One person has a data set with a sample size of 26, while the other has a sample size of 8,000. Surely rational prior model probabilities cannot ignore *n* and *K*. Clearly these issues should enter into consideration -- this is the notion of a savvy prior.

BIC is derived on the whim that each model has equal prior probability; in this case 0.333. Equal priors on models seem generally poor in this example; how could one argue for equal priors? Rather than make subjective judgments of model priors, we suggest basing the decision on sound (information) theory and avoid the entire set of "hidden difficulties" in the Bayesian approach in general. LB's appendix seems irrelevant as it relates to the impact of sample size on priors on parameters not priors on models.

GIVEN THE MODEL SET, ONE MODEL IS ALWAYS THE K-L BEST

LB: p2628 outline a hypothetical situation where "… there could be no uncertainty in the model rankings…" but "… its AIC weight need not be 100%." LB then conclude that "AIC weights cannot be interpreted as probabilities." This statement makes no sense; if there is "no model uncertainty" then that best model must have an AIC weight (model probability) of 1.0. There is also a misunderstanding here perhaps confusing a "true model" vs. a model that is closest to truth, given the data. Given a model set of size *R*, one of these models *is* the closest to truth: the K-L best model is always one of the models in the given set. The model probabilities measure the uncertainty with respect to the model *estimated* to be best (closest to truth), given the data. This is quite different from saying that one of the models *is* truth.

PREDICTION: AICc vs. BIC

In judging our faith in a model, LB: p2634 emphasize prediction is the "ultimate" rather than goodness-of-fit tests. We are aware of few attempts to fairly compare AICc vs. BIC in a predictive setting. We reported on a comparison based on predictive mean squared error (PMSE) in a linear regression setting with 13 predictor variables in a real data set pertaining to human body fat (Burnham and Anderson 2002: Section 6.4.3). These results show superiority of AICc in prediction: the PMSE values (times 10^6) were 5.685 for AICc and 7.659 for BIC, a 34% difference. In each case, model averaging provided even smaller values of the PMSE (4.854 and 5.882, respectively). More examples such as this are needed to draw a general conclusion.

Burnham and Anderson (2002:214-215) provide chain binomial examples where AICc has a smaller MSE than BIC, even when sample size was 10,000. We routinely advocate measures of prediction (when that is a goal) combined with goodness-of-fit tests, analysis of residuals, and other means to assess the model fit.

The results above suggest that Bayesian approaches are not "convincingly superior" (LB: p2627). Neither approach is always better and it misrepresents the issue if claims are made to the contrary.

MC SIMULATION NOT ALWAYS INFORMATIVE

It is important to recognize that simulation studies can be done to support either BIC or AICc; this is because they have different objectives. This makes clear interpretation difficult unless one knows how the simulations were done. MC simulation studies favor BIC if they (1) employ large sample size, (2) have a few big effects and no small or tapering effects, (3) allow the generating model to be in the model set, and (4) make final judgment concerning the worth of the criterion based on the proportion of the simulations where the generating (true) model was selected. In contrast, simulations favor AICc if (1) data are generated from a highly dynamic model with many tapering effect sizes, nonlinearities and interactions, (2) the generating model is not in the model set, (3) sample size can be small, moderate or large, and (4) final judgment is based on achieved confidence interval coverage or predictive mean squared error. Research people must decide which end of this spectrum they lean toward as they select analytical methods to understand the information real data.

Much of the research on BIC has looked at very simple models where computer truth includes a few (2-5 often) big effects while other potential effects are of no importance. This "all or nothing" view is far from reality where increasing sample size conveys the ability to find the smaller effects, nonlinearities, and interactions. BIC is limited in its performance in these real world situations.

In general the relevant literature shows that BIC tends to under-fit if sample size is small and effects are at all interesting (not just large effects or no effects). Such under-fitting results in biased parameter estimates, overestimates of precision, and achieved confidence interval coverage below the nominal level. BIC generally performed poorly in relation to AICc in the large scale Monte Carlo simulation studies presented in the book by McQuarrie and Tsai (1998).

SUMMARY AND CONCLUSIONS

1) In our perspective, the Bayesian vs. non-Bayesian focus of LB should be downplayed in several respects. BIC, AIC, and AICc can all be viewed as having Bayesian foundations; the focus should be on a science philosophy about truth and models and priors on models. In spite of its name, BIC is not derived from or related to information theory. The theory underlying model selection in a Bayesian framework is still developing and some Bayesians do not even believe in model selection (*the* model is known in their philosophy). Present approaches include DIC and the reverse jump Monte

Carlo Markov Chain and both seem problematic in the analysis of real data where some reasonable complexity is present.

2) It is important for people to realize that AIC was derived as a bias corrected estimator of expected Kullback-Leibler information, it is directly related to entropy, and its units are in information (–log(Prob)). It is the quantification of "information" itself that is one of the strengths of these new approaches.

3) Ideally, people in the statistical sciences need to reach agreement on the concept and implications of assuming that a true model exists and that it is in the model set. It is too easy for a mathematician to state "p_i is the prior probability of model i" without finishing the definition. This deliberate vagueness has not served us well in the past and will not serve us well in the future.

4) It seems unwise to attempt to claim that the Bayesian approach is the only gateway to multimodel inference. Information theory is a huge scientific endeavor and has affected nearly all aspects of our society even though it is perhaps less than 60-70 years old. While many statisticians have at least some level of understanding of Bayesian approaches, far fewer might claim the same level of understanding of information theory. Perhaps we need to be more open to new approaches.

5) LB: p2633 emphasize that caution is needed in using BIC to compute approximate posterior model probabilities. We cannot recommend the use of BIC in the analysis of complex real data.

6) All statistical methods involve various approximations and idealizations and many have some reliance on asymptotics. It should not be thought that Bayesian methods are free from these issues. Beyond that, there is 200+ year debate about the arbitrariness and

subjectivity of the Bayesian approach in scientific applications. Vague priors on model parameters have lessened this debate among scientists and we see this as encouraging. We see substantial convergence between the classic likelihood methods and the Bayesian methods and this has important ramifications for the future of the field. Still, "buried assumptions" relate to the existence of a true model being in the model set and the arbitrary model priors in the Bayesian paradigm.

7) More comparisons of predictive mean squared error would be interesting to focus on AICc, DIC, BIC, and fully Bayesian MCMC. This must probably rely on MC simulation studies where the generating model has some substantial figment of reality and its interacting complications. Such studies would shed light on the advantages of different classes of model priors. Still, doing studies of PMSE in a fully Bayesian way might not be computationally feasible for some years to come. Of course, many science problems are not just predictive.

8) It is important to use AICc instead of the asymptotic AIC; LB's: p2634 references to the "tendency of AIC to over-fit" stems from considering the wrong criterion when the ratio of sample size to the number of model parameters is low.

9) LB: p2634 suggest "… there is substantial room for improvement." We suggest two things for consideration. First, use AICc instead of AIC; both are derived in the face of parameter uncertainty. Second, would be to encourage more hard thinking about the plausible science hypotheses and (approximating) models to represent them so that the model set is supportable.

LITERATURE CITED

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. Pages 267-281 *in* B. N. Petrov and F. Csaki, (eds.) Second International

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

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: A practical information-theoretic approach. 2^{nd}. Ed., Springer-Verlag, New York.

Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference: understanding AIC and BIC in model selection. Sociological Methods & Research **33**: 261-304.

Givens, G. H., and J. A. Hoeting. 2005. Computational statistics. John Wiley and Sons, Hoboken, New Jersey, USA.

Jeffreys, H. 1973. Scientific inference. Cambridge University Press, Cambridge, UK.

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

Link, W. A., and R. J. Barker. 2006. Model weights and the foundations of multimodel Inference. Ecology 87:2626-2635.

McQuarrie, A. D. R., and C-L. Tsai. 1998. Regression and time series model selection. World Scientific Publishing Company, Singapore.

Sugiura, N. 1978. Further analysis of the data by Akaike's criterion and the finite correction. Communications in Statistics, Theory and Methods **A7**, 13-16.