ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT

Journal of Mathematical Psychology 47 (2003) 580–586

Journal of Mathematical Psychology

http://www.elsevier.com/locate/jmp

Book review

How many parameters does it take to fit an elephant?

Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. By Burnham and Anderson. 2nd Edition, Springer, New York, 2002, xxvi + 496 pp., price $79.95, ISBN 0-387-95364-7.

Reviewed by Eric-Jan Wagenmakers

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

Kenneth P. Burnham (a statistician) has applied and developed statistical theory for 30 years in several areas of life sciences, especially ecology and wildlife. He is the recipient of numerous professional awards. Dr. Burnham is a fellow of the American Statistical Association.

David R. Anderson is a senior scientist with the Biological Resources Division within the U.S. Geological Survey and a professor in the Department of Fishery and Wildlife Biology, Colorado State University. He is the recipient of numerous professional awards, including the Meritorious Service Award given by the U.S. Department of the Interior.

The reviewer, Eric-Jan Wagenmakers, received his Ph.D. in psychology under the direction of Jeroen G. W. Raaijmakers. In 2000 he received a Fulbright scholarship to work with Richard M. Shiffrin at Indiana University. From 2001 to 2003 he was a postdoctoral research fellow with Roger Ratcliff at Northwestern University. His research interests include model selection, time series analysis, and bootstrap methods.

When a psychologist makes inferences from a limited set of data, she generally uses a model to differentiate the replicable, structural information from the idiosyncratic, non-replicable information. The quality of inference thus relates directly to the quality of the model: An appropriate model will capture a lot of the structure and will at the same time treat idiosyncratic information as 'noise', thereby maximizing the probability of correct inference.

The book under review here, ''Model Selection and Multimodel Inference'' (henceforth MSMI) by Ken Burnham and David Anderson is one of the relatively few books that is entirely devoted to model selection. The book includes an in-depth discussion of the general philosophical issues involved in model selection and very clear and non-technical description of the proposed methodology. The authors illustrate their ideas with numerous examples, all taken from the field of biology. Fortunately for readers whose main interest is in psychology, the examples are either very general (e.g., variable selection in regression analysis) or they can be readily translated to psychological phenomena. For instance, models for mortality rates could be interpreted as models for human forgetting by replacing the concept of an animal with the concept of a memory trace for a studied item.

MSMI is mainly devoted to one specific model selection method, namely Akaike's Information Criterion (AIC; Akaike, 1974; Bozdogan, 1987; Parzen, Tanabe, & Kitagawa, 1998; Sakamoto, Ishiguro, & Kitagawa, 1986). A substantial part of the book shows how the AIC can be used for multimodel inference, a theme that was hitherto mostly associated with Bayesian model averaging (cf. Wasserman, 2000).

To foreshadow the conclusion, I believe MSMI is a very provocative, well-written book. It will certainly be very useful to psychologists who want to know about the philosophical and statistical foundations of AIC. The practical examples from the book invite the reader to apply the proposed methodology in his or her own research. The first five chapters require only a rudimentary knowledge in statistics, making the book perfectly suitable for an introductory course on model selection. The authors have taken considerable care to motivate their claim that AIC is the best general purpose method for model selection in the life sciences, using both analytical derivations and simulations. I believe some of the work presented in MSMI could be the starting point of a exciting scientific discussion.

Below I will first summarize several key ideas advocated by the authors of MSMI, and then turn to a more evaluative discussion of the book. Please note that this review is concerned with the second, much

doi:10.1016/S0022-2496(03)00064-6

improved edition of MSMI. The second edition contains more material and fewer mistakes than the 1998 edition. Also, the presentation of ideas is much more structured in the second edition than it was in the first.

1. Selective overview of key issues

Chapters 1 and 2 of MSMI provide the philosophical background and basic ideas. Chapters 3–5 illustrate the proposed methodology using many examples from biological research. Finally, Chapters 6 and 7 are mostly concerned with statistical theory. As a whole, MSMI aims to present a coherent and principled strategy for the analysis of empirical data, with an emphasis on AIC model selection and multimodel inference.

The motivation for model selection is ultimately derived from the *principle of parsimony* (cf. Forster, 2000). In the glossary, Burnham and Anderson define parsimony as

> The concept that a model should be as simple as possible concerning the included variables, model structure, and number of parameters. Parsimony is a desired characteristic of a model used for inference, and it is usually defined by a suitable tradeoff between squared bias and variance of parameter estimators. Parsimony lies between the evils of under- and over-fitting.

The principle of parsimony is also known as Occam's razor: to remove all that is unnecessary ("it is vain to do with more what can be done with fewer", William of Occam, from Grünwald, 2000, p. 133). A model that is not very parsimonious will capture relatively much of the idiosyncratic information (i.e., noise) in the data. Note that by just adding parameters, it is possible to fit almost anything (about 30 parameters suffices when fitting an elephant, Burnham & Anderson, 2002, p. 30). Such a model might provide an excellent fit to the data at hand, but because its parameter estimates are quite variable the model will generalize poorly to novel data sets. Hence, inference from unparsimonious, over-fitted models is hazardous and should be avoided. Of course, a model that captures relatively little structural information (i.e., an under-fitted model) is also not well suited for inference. Ideally, inference should be based on simple models that describe the data well. The authors argue that AIC obeys the principle of parsimony as a by-product of its derivation, to which we turn next.

One of the core ideas in MSMI is that in the life sciences, an exact understanding of reality is an unattainable goal. Thus, truth is effectively considered to be infinitely dimensional, and we can only hope to find a useful approximation to this complex, unknown truth. In the following, f and g will denote continuous probability distributions that refer to unknown truth and an approximating model, respectively. From information theory we know that the distance between f and g is given by the Kullback–Leibler information I (e.g., Kullback & Leibler, 1951):

$$I(f,g) = \int f(x) \log \left(\frac{f(x)}{g(x|\theta)} \right) dx, \tag{1}$$

where x denotes the data and θ is a vector of free parameters. $I(f,g)$ is often called K–L distance and can be interpreted as the information lost when g is used to approximate f. The K–L distance $I(f,g)$ is zero when $g=f$, and positive otherwise. Using statistical expectations with respect to truth f instead of integrals, Eq. (1) can be re-written as

$$I(f,g) = E_f[\log (f(x))] - E_f[\log (g(x|\theta))]. \tag{2}$$

Since truth is the same for all candidate models, it drops out as a constant C, and hence $-E_f[\log (g(x|\theta))] = I(f,g) - C$ gives the *relative* K–L distance. Note that the parameter values for g are not known, but have to be estimated. The maximum likelihood parameter estimates for $\hat{\theta}$, based on some specific data set x, will generally not equal their true values, which can ideally be approximated by the average of $\hat{\theta}$ over replicate data sets from the same data generating process. This parameter uncertainty can be taken into account by using an approach similar to cross-validation. Denoting a replicate data set by y, the minimum expected K–L distance is $E_y[I(f,g(\cdot|\hat{\theta}(y)))]$, which is larger than the K–L distance based on the unknown, ideal parameter estimates θ_o. After some rewriting, the minimum expected K–L distance is given by

$$E_y E_x[\log (g(x|\hat{\theta}(y)))], \tag{3}$$

where x and y are independent samples from the same data generating process, and expectation is with respect to unknown truth. Asymptotically, this quantity can be estimated by the log likelihood minus the number of parameters (Section 7.2 in MSMI gives the mathematical details). Multiplied by –2, this quantity yields the AIC:

$$AIC = -2 \log L + 2K, \tag{4}$$

where L is the maximum log likelihood and K is the number of free parameters. The model with the smallest AIC value thus has the smallest expected K–L distance and is the closest approximation to the complex truth. Thus, AIC rewards models for goodness-of-fit through $-2 \log L$, and punishes models for lack of parsimony through the penalty term $2K$ that solely depends on the number of free parameters. When the number of free parameters is relatively large compared to sample size, the authors strongly recommend a small-sample version of AIC (e.g., Hurvich & Tsai, 1989):

$$AIC_c = -2 \log L + 2K + \frac{2K(K+1)}{n-K-1}, \tag{5}$$

where n is sample size.

The original derivation of the AIC assumed that the data generating process is among the set of candidate models. Later, a more general derivation resulted in a different information criterion, TIC (Takeuchi's Information Criterion; Takeuchi, 1976). TIC does not assume the candidate model is in the set, and this is more in line with the research philosophy advocated by the authors. However, MSMI argues that the calculation of TIC is more involved than calculation of AIC, and that TIC leads to more variable results. Further, simulations show that the penalty term of AIC is very close to the penalty term of TIC, especially when a model gives a reasonable fit. For models that do not fit the data, the precise form of the penalty term for lack of parsimony will be much less important than the penalty term for lack of descriptive accuracy. For these reasons, the authors prefer AIC over TIC.

When AIC model selection is used in practice, the aim is almost invariably to select a single best model and base all inferences on that model. This practice might stem from the common misconception that any difference in AIC, no matter how small, is somehow meaningful, so that only the AIC best model is relevant. Burnham and Anderson outline a straightforward method on how to assess differences in AIC (cf. Wagenmakers & Farrell, in press). First, for each candidate model i, one calculates the difference in AIC between it and the AIC-best candidate model: $\Delta_i = AIC_i - AIC_{\text{best}}$. This highlights the fact that only differences in AIC are important, not the absolute values. The likelihood of model i, given the data, is proportional to $\exp\left(-\frac{1}{2}\Delta_i\right)$. The probability of model i being the K–L best model, given the data and the set of candidate models, is then given by the Akaike weight

$$w_i = \frac{\exp\left(-\frac{1}{2}\,\Delta_i\right)}{\sum_{r=1}^{R}\,\exp\left(-\frac{1}{2}\,\Delta_r\right)}, \tag{6}$$

where R is the number of candidate models. Thus, if $R = 2$ and $w_1 = 0.6$, this means that the second best model is still very likely to be the K–L best model (i.e., $w_2 = 0.4$). In this case, the likelihood ratio in favor of the AIC best model is only 1.5.

Chapter 4 of MSMI discusses a further purpose of the model weights from Eq. (6): multimodel inference. When model selection is used only to select the best model, then inference is effectively conditional on that model. Hence, the uncertainty that is associated with the model selection enterprise is ignored, which leads to an overestimate of precision. In other words, it is risky to base inference on a single model when this model is not clearly superior to its competitors. Burnham and Anderson propose to eliminate this risk by basing inference on all candidates models simultaneously, weighing their impact by the Akaike weights. Thus, when θ is either a predicted value of interest, or a parameter of interest that is common to all candidate models (such as in variable subset selection in regression analysis), the model averaged value $\hat{\bar{\theta}}$ is given by

$$\hat{\bar{\theta}} = \sum_{i=1}^{R} w_i \hat{\theta}_i, \tag{7}$$

where R notes the total number of candidate models.[1]

The sampling variance of a parameter conveys information about its precision. The estimates of sampling variance that are usually reported are conditional on a specific model. An estimate of parameter sampling variance v that is *unconditional* on a specific candidate model is given by

$$\hat{v}(\hat{\bar{\theta}}) = \left[\sum_{i=1}^{R} w_i\,\sqrt{\hat{v}(\hat{\theta}_i|g_i) + (\hat{\theta}_i - \hat{\bar{\theta}})^2}\right]^2, \tag{8}$$

where g_i denotes candidate model i. This estimate incorporates the conditional variance estimate through $\hat{v}(\hat{\theta}_i|g_i)$, and a variance component for model selection uncertainty through $(\hat{\theta}_i - \hat{\bar{\theta}})^2$. Monte Carlo simulations show that the unconditional variance is a more accurate reflection of precision than is the conditional variance (which is biased downward). Multimodel inference generally results in predictions and estimates that have less bias and more precision. The practical advantages of multimodel inference are illustrated throughout Chapter 4.

Chapters 5 and 6 deal with a wide variety of model selection issues. In particular, the authors compare performance of AIC model selection to that of model selection using the Bayesian Information Criterion (BIC; Schwarz, 1978). BIC is given by $-2\log L + K\log n$, where n is the number of observations (for details see Kass & Raftery, 1995). BIC is often reported to be superior to AIC, in particular because as $n \to \infty$, BIC but not AIC will select the correct data generating model with probability $p \to 1$. Hence, it is often said that BIC is 'dimension consistent' (Bozdogan, 1987). For $n > e^2 \approx 8$, the BIC penalty term is more strict than the AIC penalty term, thus resulting in the selection of relatively low-dimensional models. Burnham and Anderson perform several Monte Carlo simulations, and from these they conclude that "BIC selection cannot be recommended. It requires very large sample sizes to achieve consistency; and typically, BIC results in a selected model that is underfit (e.g., biased parameter estimates, overestimates of precision, and achieved confidence interval coverage below that achieved by AIC$_c$-selected models)." (p. 213).

This is a surprising conclusion, since many Monte Carlo studies have shown BIC model selection to

[1] When θ only occurs in a subset of candidate models, one can either base inference about θ on that specific subset, or one can set $\hat{\theta}$ to zero for those candidate models that do not contain it.

perform better than AIC model selection. However, these previous Monte Carlo studies can generally be characterized as follows: (1) there is a true model, and it is in the set of candidate models; (2) the true model has relatively few parameters, and (3) performance is evaluated by how often the true data-generating model is recovered. Burnham and Anderson argue that such simulations are not realistic. In many of the Monte Carlo simulations from MSMI, the true data-generating model is not in the set of candidate models, and, more importantly, the data-generating model has *tapering* effects such that, for instance, treatment effects gradually diminish over time. This means that as the number of observations increases, evermore small (but real) effects can be uncovered. Tapering effects are consistent with the philosophy of an infinitely dimensional truth, and, it is argued, they are more representative of biological reality than models with fixed effects.

In the MSMI simulations, performance is assessed not by correcting model recovery rates—this can only be done when the correct model is in the set of candidate models—but instead by achieved confidence interval coverage (i.e., whether the confidence interval for a given parameter estimate encompasses the true parameter value).

Chapter 6 also compares AIC model selection for nested models to the likelihood ratio test (LRT). The LRT is based on the fact that minus two times the difference in log likelihood is χ^2 distributed with the number of degrees of freedom equal to the difference in the number of free parameters. The authors point out that AIC and LRT are based on quite different procedures. AIC is based on model selection (i.e., minimizing the expected K–L distance), and LRT is based on a null-hypothesis testing framework. In practical applications, one often finds that LRT is used for nested models, whereas AIC is reserved for nonnested models only—conceptually, this is an awkward mixture of analysis paradigms.

Burnham and Anderson argue that model selection based on LRT rather than AIC does not have a "sound theoretical basis". The authors provide the following scenario to illustrate their claim. Assume a set of nested models, each successive model having one additional free parameter. Also, assume the AIC values for each of the models are exactly the same; hence, the data supports every model to an equal extent. When LRT is applied to the above situation, it turns out that the null-hypothesis of model g_i, where the subscript denotes the number of free parameters, is rejected with increasing strength as the number of additional parameters in the alternative model increases. For instance, the difference between g_i and g_{i+1} is $\chi^2_{df=1}$ distributed. For a difference in $-2\log(L_i/L_{i+1})$ of 2 (i.e., equal AIC values), the p value is 0.157. When model g_i is compared to model g_{i+30}, however, this results in a p value of 0.001. The authors cite Akaike (1974), who explains that "The use of a fixed level of significance for the comparison of models with various numbers of parameters is wrong, because it does not take into account the increase of the variability of the estimates when the number of parameters is increased." (MSMI, p. 339).

Chapter 7 contains 'statistical theory and numerical results'. This chapter can be skipped by the reader who just wants to know how to apply and interpret AIC and multimodel inference in his own work. Section 2 is the most important section in this chapter, as it gives a general derivation of AIC. Finally, Chapter 8 provides an overall summary of the contents of the book.

2. Evaluation and discussion

2.1. Style of presentation

One feature of MSMI that sets it apart from most other books is the style of presentation. The authors have done an excellent job clarifying their procedures, thus making the material accessible to the applied biologist/psychologist/econometrician. The flip side of this is that the same ideas are repeated many times throughout the book, almost as if the authors are attempting to brainwash or indoctrinate the reader. The introduction of MSMI does contain a subtle warning for the reader, as the authors expressed hope that "(…) the text does not appear too dogmatic or idealized." (p. x). Such a statement will lead any psychologist to expect a great number of dogmas in the text. Here is a listing of the most important ones.

Dogma 1: Thou shall not commit data dredging. Instead, appropriate candidate models should be developed beforehand. A distinction should be made between exploratory data analysis (i.e., post-diction) and confirmatory data analysis (i.e., prediction). More generally, careful *thinking* should motivate the proposed candidate models. This dogma is very important—it highlights that it can be dangerous to apply model selection tools in a completely automated fashion, disregarding issues such as plausibility, explanatory adequacy, internal consistency, and interpretability (e.g., Jacobs & Grainger, 1994). Model selection is a multifaceted problem, and model selection methods address only a subset of relevant criteria such as descriptive accuracy, generalizability, and complexity (cf. Pitt, Myung, & Zhang, 2002).

Dogma 2: Thou shall not commit null-hypothesis testing. Instead, one should minimize the expected K–L distance to obtain an optimal balance between underfitting and overfitting.

Dogma 3: Truth is infinitely dimensional and can only be *approximated* by our models, but never captured completely.

Dogma 4: Thou shall conduct Monte Carlo simulations that include tapering effects. The data-generating model should preferably not be in the set of candidate models.

Dogma 5: Inference shall be based on more than one model, especially if the Akaike weight for the best model is smaller than 0.9.

Dogma 6: Thou shall not use BIC. Thou shall use AIC.

2.2. Lay-out

On the positive side, MSMI contains many helpful figures that nicely support the text. Photos of Fisher, Boltzmann, Akaike, Kullback, Leibler, Shibata, and Takeuchi further enliven the presented material. The references are remarkably up-to-date and comprehensive. On the negative side, the book contains a fair number of typographical errors, and the index is too small and selective to be very helpful (e.g., it contains the word "elephant" but does not contain the word "prior"). The problems this causes in quickly finding relevant material is exacerbated by the absence of an author index.

2.3. Contents

MSMI contains an abundance of interesting ideas, many of which can be considered open to further research (see in particular Chapter 6). Here I have selected for further discussion two important issues that are recurrent themes throughout the book. The first issue concerns the use of the bootstrap to approximate model weights, and the second issue involves the comparison to other methods of model selection, particularly BMS and MDL.

2.3.1. Bootstrap model weights

As an alternative to the direct calculation of model weights according to Eq. (6), Burnham and Anderson suggested the use of the *nonparametric bootstrap* to approximate these model weights (cf. Buckland, Burnham, & Augustin, 1997). I believe this procedure is biased, as will be made more explicit below.

The non-parametric bootstrap was introduced by Efron (1979) and Efron and Tibshirani (1993) and it is most often used to approximate the standard error for a parameter estimate (for an application of the bootstrap methodology to psychometric functions see Wichmann & Hill, 2001). Let X_n be a set of n observations, $X_n = \{x_1, x_2, \ldots, x_n\}$. A statistic such as the median or correlation coefficient is calculated based on X_n. When we take X_n to represent the population, we can sample, with replacement, from X_n to obtain new, bootstrap, samples X_n^*. The statistic of interest is then calculated for each bootstrap sample, and the standard deviation of the statistic based on its distribution over the bootstrap samples approximates the standard error.

Burnham and Anderson propose, albeit tentatively, to employ the bootstrap method to calculate model weights. That is, each bootstrap sample X_n^* is analyzed as if it were the actual data X_n: each of the candidate models is fit to the bootstrap sample, followed by the calculation of the corresponding AIC values. Based on a total of M bootstrap samples (M is usually 10,000 in MSMI), a bootstrap model weight can be derived by averaging the Akaike weights over all bootstrap samples: $\bar{w}_i^* = (1/M)\sum_{j=1}^{M} w_i^*(j)$, where i indexes the model and j the bootstrap sample. An alternative procedure is to tally the number of bootstrap samples for which a candidate model has the lowest AIC value—the model weight is then simply the proportion of bootstrap samples in which the model under consideration "wins". The problem with using these procedures, however, is that the naive non-parametric bootstrap of the AIC is biased. Bollen and Stine (1992) have shown that when the non-parametric bootstrap is used, the distribution of minus two times the likelihood ratio is no longer χ^2 distributed with degrees of freedom equal to the difference in the number of free parameters (see also Wagenmakers, Farrell, & Ratcliff, in press).

This bias occurs because although the simple model might be true for the population, it will not hold exactly for a particular sample. Thus, the bootstrap method for determining model weights is not to be recommended, unless some bias-correcting measures are taken first. To be fair, Burnham and Anderson do state that the bootstrap method occasionally fails, and they generally prefer the calculation of Akaike weights according to Eq. (6).

2.3.2. Comparison to other methods of model selection

MMSI is focused on selecting models based on the expected minimum K–L distance as estimated by AIC. As noted above, the AIC aims to correct the maximum likelihood estimation-bias that exists because the same data are used both for parameter estimation and estimation of the expected likelihood. Instead of using Akaike's asymptotic approximation for this bias (cf. Eq. (4)), it is also possible to use the bootstrap to estimate the size of the bias. Such a bootstrap extension of AIC, called EIC, has received some attention recently (e.g., Ishiguro, Sakamoto, & Kitagawa, 1997; Konishi & Kitagawa, 1996; Pan, 1999), and simulations show that EIC is especially helpful in situations of small sample size. I believe a subsequent edition of MSMI would do well to include a more in-depth discussion of EIC—the current edition devotes not even an entire page (p. 374) to this interesting procedure.

The only other method for model selection (except TIC, which is closely related to AIC, see above) that is discussed in detail is BIC. The authors show that when a specific non-uniform prior is used, Akaike weights can be interpreted as Bayesian posterior model probabilities (pp. 302–305) as approximated by BIC (but see Kass & Raftery, 1995). However, the derivation of BIC is based on large sample asymptotics. A full-fledged Bayesian approach to model selection (e.g., Kass & Raftery, 1995; Myung & Pitt, 1997; Wasserman, 2000) selects the model that has the highest posterior probability of having generated the data.

The practical disadvantage of such Bayesian model selection (BMS) is that it is relatively complex and often requires high-dimensional integrals to be approximated by computer-intensive Markov chain Monte Carlo techniques. The advantage of BMS is that it takes not only the number of parameters into account, but also their functional complexity. Recall that Burnham and Anderson defined parsimony as "The concept that a model should be as simple as possible concerning the included variables, *model structure*, and number of parameters" [italics added]. The penalty terms of AIC and BIC only take the number of parameters into account, not their functional complexity. For example, Stevens' law of psychophysics can handle both decelerating and accelerating functions, whereas Fechner's law can only account for decelerating functions (cf. Pitt et al., 2002). Thus, Stevens' law is a priori more flexible, despite the fact that it has just as many parameters as has Fechner's law.

Another important model selection philosophy that MSMI mostly ignores is *minimum description length* (MDL; e.g., Grünwald, 1998, 2000; Li & Vitanyi, 1997; Pitt et al., 2002; Rissanen, 1996, 2001). MDL implements the principle that the best model yields the highest compression of the data. Because probability relates to code length (high probabilities being associated with short codes), MDL model selection is related to (but not identical to) BMS (cf. Grünwald, 2000). The issue of structural complexity is particularly pressing when the candidate models are non-nested. For example, Burnham and Anderson (p. 15) list nine models for avian species-accumulation (from Flather, 1996), the first four being given by $E(y) = ax^b, E(y) = a + b \log x, E(y) = a(x/(b + x))$, and $E(y) = a(1 - e^{-bx})$. All these models have three parameters, but some of these models will be more complex than others. Thus, according to BMS and MDL, the question is not just *how many* parameters it takes to fit an elephant, but rather how many and *what kind* of parameters. A nice overview of different model selection methods can be found in the recent special issue of *Journal of Mathematical Psychology* (Myung, Forster, & Browne, 2000).

3. Overall evaluation

MSMI is a very useful and thought-provoking book on the advantages and practical application of AIC model selection. It is a good book to use for a interactive graduate course on model selection. The authors do not hesitate to make strong claims that invite further reading and additional research.

References

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

Bollen, K. A., & Stine, R. (1992). Bootstrapping goodness of fit measures in structural equation models. *Sociological Methods and Research*, *21*, 205–229.

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

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, *53*, 603–618.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Flather, C. H. (1996). Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *Journal of Biogeography*, *23*, 155–168.

Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology*, *44*, 205–231.

Grünwald, P. (1998). *The MDL principle and reasoning under uncertainty*. Ph.D. thesis, University of Amsterdam, available as ILLC Dissertation Series, DS 1998-03.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, *44*, 133–152.

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

Ishiguro, M., Sakamoto, Y., & Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Annals of the Institute of Statistical Mathematics*, *49*, 411–434.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition—sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, *29*, 1311–1334.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Konishi, S., & Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, *83*, 875–890.

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

Li, M., & Vitanyi, P. (1997). *An introduction to Kolmogorov complexity and its applications* (2nd ed.). New York: Springer.

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). Model selection (Special issue). *Journal of Mathematical Psychology*, *44*(1–2).

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Pan, W. (1999). Bootstrapping likelihood for model selection with small samples. *Journal of Computational and Graphical Statistics*, *8*, 687–698.

Parzen, E., Tanabe, K., & Kitagawa, G. (1998). *Selected papers of Hirotugu Akaike*. New York: Springer.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

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

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, *47*, 1712–1717.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike information criterion statistics*. Dordrecht: Reidel.

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

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

Wagenmakers, E.-J., & Farrell, S. (in press). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*.

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (in press). Naïve nonparametric bootstrap model weights are biased. *Biometrics*.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*, 92–107.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics*, *63*, 1314–1329.

Eric-Jan Wagenmakers
Department of Developmental Psychology, Room 1001
University of Amsterdam, Roetersstraat 15
1018 WB Amsterdam, The Netherlands
E-mail address: ej@northwestern.edu