PSYCHOMETRIKA—VOL. 54, NO. 3, 539–541
SEPTEMBER 1989
REVIEWS

REVIEWS

Y. Sakamoto, M. Ishiguro, G. Kitagawa. *Akaike Information Criterion Statistics*. Dordrecht/Boston/Lancaster/Tokyo: D. Reidel, 1986.

Estimation of the unknown parameters in a parametric probability model is one of the basic problems of statistics. It has been very thoroughly studied, both in the finite sample case and in the large sample case. But the theory of estimation requires that the model be completely specified, up to the values of the unknown parameters. And in many situations in the applied sciences we do not know the precise form of the model, we only suspect that it is in some given class of models. Even if we restrict our choice of examples to psychometrics, we can readily find hundreds of them. We are prepared to assume that the common factor analysis model is true, but we do not know the precise number of factors. We are willing to assume that the regression is polynomial, but we do not know the degree. An analysis of variance model or loglinear model obtains, but we are not sure about the interactions we want to include. And so on. In cases such as these we do not only have to compute point estimators of the unknown parameters, but we also have to select a model from the class of available models. Until recently this problem was usually solved by ad hoc search techniques. The elbow-criteria of factor analysis, the various rules to test interactions in the analysis of variance, and the variable selection methods in stepwise regression are good examples of such techniques. Formal statistical analysis, based on decision theory, turned out to be "horrendously difficult" (Anderson, 1962). The much weaker form of simultaneous statistical inference was constructed especially for this purpose.

Recently the combination of model selection with point estimation has become even more common. Sophisticated computer programs such as LISREL require choices from a very large class of possibly very complicated models. This makes the "horrendously difficult" problems studied by Anderson easy by comparison. Because there is so little prior knowledge in the areas in which LISREL is typically applied, there has been a tendency to automate the model selection process. LISREL computes modification indices to guide this process, and comparable programs such as EQS or COSAN have similar search strategies built in. Thus, in an admittedly extreme case, the user of the program inputs a covariance matrix and the program outputs a model. *Deus ex machina*. Instant science. Moreover the free parameters of the selected model are estimated efficiently, and confidence interval information is also provided. Although the result of applying computer programs such as these can look very impressive indeed, some people are not impressed by the general procedure (Freedman, 1987).

One of the more popular criteria that can help in a computerized model selection process is Akaike's information criterion (AIC). The September 1987 issue of *Psychometrika* discusses the AIC and its applications in some detail, and the book reviewed here is the first comprehensive work on theory and applications of the AIC. It is the translation of a book published in Japan in 1983. In our rather long introduction we have tried to place AIC in a historical and methodological context, and we have tried to indicate that serious dangers are connected with its uncritical use. It will be interesting to evaluate this book in the context of this discussion. But first we shall review its factual contents.

The book starts out with a short introduction to probability and statistics, rapidly covering random variables, probability distributions and discrete and continuous models. This introduction is probably intended to make the book relatively self-contained, so that it can be used as a course in statistical methods based on the AIC, for instance for engineers and people working in the theory of systems (this is also suggested by the various forewords and prefaces in the book). If the book is really used in such a way, that would be very unfortunate. There is much more to statistics than the AIC.

We now come to the most interesting part of the book, which introduces the AIC. First the problem of estimating the unknown probability distribution generating the data is introduced. For this purpose we need a distance-like measure to compare the probability distributions in the statistical model to the true distribution, and the authors choose, without much argumentation, the Kullback-Leibler information quantity. This is, in the discrete case, $\Delta(\pi, p) = 2\Sigma\pi_i \log(\pi_i/p_i)$, with $\pi = \{\pi_i\}$ the true distribution and $p = \{p_i\}$ the model. If we are not dealing with a simple model p but with a set of models $\mathcal{P}$, then we use $\Delta(\pi, \mathcal{P}) = \min\{\Delta(\pi, p) | p \varepsilon \mathcal{P}\}$ to indicate the quality of the model. In order to compute $\Delta(\pi, \mathcal{P})$ we must project π on the model $\mathcal{P}$, using the metric Δ.

In practice we do not know π, but we observe a vector of proportions $\hat{p}$. The projection of $\hat{p}$ on the model is the maximum likelihood estimate $\pi_{\mathcal{P}}(\hat{p})$. Thus $\Delta(\hat{p}, \mathcal{P}) = \Delta(\hat{p}, \pi_{\mathcal{P}}(\hat{p}))$. In the usual statistical theory we use the fact that $n\Delta(\hat{p}, \mathcal{P})$ has asymptotically a central chi square distribution if the model is true, that is, if $\pi \varepsilon \mathcal{P}$. We could compare models $\mathcal{P}_1, \ldots, \mathcal{P}_r$ by comparing the values of $\Delta(\hat{p}, \mathcal{P}_s)$, but this has the disadvantage that the more general model is always better, and the saturated model is always best. Akaike proposes to use a replication of $\hat{p}$, that is, a random vector $\hat{q}$ which is independent of $\hat{p}$ and has the same distribution. We try to eliminate the effect of chance capitalization by computing the maximum likelihood estimate from $\hat{p}$, and then comparing it with $\hat{q}$. This is the basic idea of cross validation, in a modern disguise. Define the mean expected log likelihood as $\mathbf{E}_p\mathbf{E}_q\{n\Delta(\hat{q}, \pi_{\mathcal{P}}(\hat{p}))\}$. This is the basic quantity used to compare models. It is difficult to compute or estimate directly, except by cross validation methods. There is a convenient estimate available however, which is $n\Delta(\hat{p}, \mathcal{P}) - \text{df}$, with df the number of degrees of freedom of the model. This is the AIC statistic, which is minimized over models. Large models are penalized because they have a small number of degrees of freedom.

The book explains the above theory in a less than satisfactory way. In many cases the fact that Akaike has proposed a certain quantity seems to be sufficient justification for the authors to adopt it. This gives the book a somewhat religious flavor. On the other hand the mathematics is clear, much clearer than the exposition of Bozdogan (1987) in the AIC issue of *Psychometrika*. It is curious that most AIC people, including Akaike, seem to have difficulty explaining their principles in simple terms. The rather elementary mathematics get interwoven with unnecessary technicalities, offending Bayesianisms, the likelihood principle, excursions into the philosophy of science or even into thermodynamics, and so on.

In the next 100 pages of the book the AIC principle is applied to regression, analysis of variance, contingency table analysis, and multivariate multinormal analysis. This is not very spectacular, but there are nice examples, graphs and curves in these chapters, and one gets a clear feeling on how the AIC works. This part of the book can be read as a short course in statistics based on maximum likelihood methods, which are even more prominent in the AIC approach to statistics then in the usual approach. The new thing is that usually many models are considered for the same data structure, and these models are compared by using the AIC. The gain seems to be in the systematic exploration of more models, not in the mechanical procedure of choosing the one with

the smallest AIC. The basic methodological message is of course trivial: if one has a criterion to rank models, then one can use that criterion to rank models.

The final 50 pages consist of unappealing FORTRAN programs, with output. No doubt the programs are useful, but copying code into a book seems a rip-off. If the publisher wants to do something useful, he should include a floppy with a code, or give an address where such a floppy can be obtained, if necessary at some additional cost. Using the ugly output of a line printer to fill a substantial part of a book, which costs as much money as a small automobile, seems an insult.

Let us summarize our conclusions. The theoretical part of the book is useful, although not complete and not without flaws. The practical applications are nicely done, and show the workings of the techniques. The basic methodology behind the AIC has an unsympathetic mechanical aftertaste. There are many ways to rank models, and the pluralism of allowing a lot of models should be extended by allowing a lot of criteria to rank them as well. We are forever seeking certainty in our pursuit of the truth. The AIC can help, if used sensibly. It does not come with a guarantee. And model selection problems are still horrendously difficult.

UNIVERSITY OF CALIFORNIA AT LOS ANGELES *Jan de Leeuw*

References

Anderson, T. W. (1962). The choice of the degree of a polynomial regression as a multiple decision problem. *Annals of Mathematical Statistics*, *33*, 255–265.

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

Freedman, D. A. (1987). As others see us: A case study in path analysis. *Journal of Educational Statistics*, *12*, 101–128.