

AIC and Large Samples

I. A. Kieseppä^{†‡}

I discuss the behavior of the Akaike Information Criterion in the limit when the sample size grows. I show the falsity of the claim made recently by Stanley Mulaik in *Philosophy of Science* that *AIC* would not distinguish between saturated and other correct factor analytic models in this limit. I explain the meaning and demonstrate the validity of the familiar, more moderate criticism that *AIC* is not a consistent estimator of the number of parameters of the smallest correct model. I also give a short explanation why this feature of *AIC* is compatible with the motives for using it.

1. Introduction. It is well-known that the Akaike Information Criterion (*AIC*), a model choice criterion in which philosophers of science have shown a considerable amount of interest during the last few years, does not produce *asymptotically consistent* estimates for the number of parameters of the correct model (see, e.g., Woodroffe 1982, 1182). This criticism is concerned with what happens in the limit when the available sample becomes larger, and when one uses *AIC* for choosing between some fixed set of statistical models on the basis of the sample.

When *AIC* is applied to the models, the number of the parameters of the model that it leads one to choose can be viewed as an estimate of the number of parameters of the smallest correct model. This is particularly obvious in the context of the curve-fitting problems to which philosophers have until now dedicated most of their attention. In these problems one makes a choice between the nested models $M_{pol-0}, M_{pol-1}, M_{pol-2}, \dots$ which are such that each model M_{pol-k} ($k = 0, 1, 2, \dots$) is a model with $k + 1$ parameters, and contains all the curves of at most k th degree.¹ In this case

[†]To contact the author, write to Department of Philosophy, P. O. Box 9 (Siltavuorenpenger 20 A), 00014 University of Helsinki, Finland; e-mail: i.a.kieseppa@helsinki.fi.

[‡]I would like to express my gratitude to Stanley Mulaik and Malcolm Forster for our discussions on the topics addressed in this paper.

1. To be precise, the practice of viewing models of as collections of curves is legitimate when the error distribution, which determines the extent to which the observed values are displaced from the correct curve, is assumed to be known.

the smallest model M_{pol-0} is a one-parameter model which contains all horizontal straight lines, the model M_{pol-1} is a two-parameter model which contains all straight lines, and the model M_{pol-2} is a three-parameter model which contains all curves of at most second degree (i.e., straight lines and parabolas). In this case the number of the parameters of the chosen model is also an estimate for the degree of the correct curve: one can view it as an answer to the question whether the correct curve is a straight line, a parabola, or a curve of some more complicated shape.

By definition, a *consistent* estimator of an unknown quantity is an estimator which is such that its value *converges stochastically* to the actual value of the quantity as the sample size grows larger (see, e.g., Hogg and Craig 1965, 246). When this definition is applied to estimators of the number of parameters of the smallest correct model, it means that, as the sample size grows, the probability that the estimate is correct approaches one. If one uses a finite sample for choosing between statistical models one of which is actually correct, one cannot usually know for sure that one has picked the right one. However, if a model choice criterion is asymptotically consistent as an estimator of the number of parameters, one can at least know that in this case the probability of using a model which has the wrong number of parameters will approach zero as the sample size grows. However, AIC is not consistent in this sense.

Stanley A. Mulaik has recently addressed a variety of topics which are related to the statistical model selection criteria in an interesting paper (Mulaik 2001). Some of his arguments are concerned with the behavior of AIC in the limit in which the sample size approaches infinity. On the basis of a mathematically incorrect argument, Mulaik makes a claim which is much more radical than the familiar criticism that AIC is not consistent in the sense which was explained above. He claims that “in the limit” AIC does not distinguish between a saturated model (i.e., a model which has so many adjustable parameters that it can be made to fit the evidence perfectly independently of what the evidence is like) and a smaller, correct model (231). According to Mulaik, this “undermines the use of the *AIC* in attempts to explain the role of parsimony in curve-fitting and model selection” (231).

Below I shall first present the results to which Mulaik appeals in his mathematically incorrect argument. I shall show that these results do not imply that, in the limit of large samples, AIC could not be used for discarding models with many parameters when a model with a smaller number of parameters is correct. I shall also illustrate the correctness of the closely related, more moderate criticism of AIC according to which it is not consistent. In order to give a clear presentation of Mulaik’s argument it will be necessary to begin by considering on a more general level the

model choice problems which he discusses. These problems belong to the field of *factor analysis*.

2. Some Background. A factor analytic model is concerned with the connections between some observed variables and a number of other variables whose values have not been observed. In the typical case which Mulaik uses as his example the available measurement results consist of the measurements of n different quantities for each item in the available sample, which is a sample of size N . When these measurement results have become available, one can use them for calculating the *sample variance* of each of the n measured quantities, and the *sample covariance* between each pair of different measured quantities. Since there are $(1/2)n(n-1)$ such pairs, the number of variances and covariances which can be calculated in this manner is $n + (1/2)n(n-1) = (1/2)n(n+1)$. Among other things, factor analysis provides models for the values of such variances and covariances.²

The values of such variances and covariances are regularly represented in the form of a table called the *covariance matrix*. A factor analytic model postulates that, besides the observed variables, there are also unobserved *latent variables*, and specifies connections between them and the observed variables. These connections are expressed by probabilistic equations, in which the value of each observed variable is equated with a sum which contains the values of the unobserved variables and an error term. Normally, these equations also contain adjustable parameters which are such that, if one gives some particular, fixed values to them and to the error terms, the model will yield a value for each item in the covariance matrix. This implies that, when the values of the parameters have been specified, the model will yield for the observed covariance matrix a probability distribution which depends on the probability distributions of the error terms. When one further sets the error terms to zero, the model will predict which values will appear in the observed covariance matrix.

The values of the parameters of a model can be estimated by choosing for them the values which have the maximal likelihood relative to the observed covariance matrix. Just like in the case of curve-fitting problems, in which the models with many parameters will, in general, contain curves which are close to the points which represent observations, the factor analytic models with many parameters will, in general, have a larger likelihood than the models with few parameters. In the extreme case in

2. The factor analytic models which I use here as my illustration, and to which also Mulaik seems to be referring, are discussed at length in e.g. Bollen (1989), 226–318. For a broader introduction to the relevant philosophical and statistical literature, see Kiesepää 2001, Sakamoto et al. 1986.

which there are just as many parameters—i.e., $(1/2)n(n+1)$ —in the model as there are variance and covariance values which the model is supposed to explain, the equations which connect these values with the parameters will have a solution in which all the error terms have the value zero. In this case and the prediction concerning the covariance matrix that the model yields will be identical with the covariance matrix which has been observed.

A model like this is a *saturated* model. Also more generally, a model is called saturated if it has so many parameters that it will fit the evidence perfectly, no matter what the evidence is like. In his argument Mulaik appeals to a result which is concerned with the likelihood ratio of a saturated model and the model which is under consideration (Mulaik 2001, 230). The references which he gives to this result are McDonald 1989 and McDonald and Marsh 1990, but it seems that a clearer presentation of the result to which he appeals has been given in e.g. Bozdogan 1987, a paper to which McDonald and Marsh also refer (1990, 251).

Bozdogan contrasts a saturated model with K parameters with a smaller model with k parameters, and considers a likelihood ratio statistic with which the success of the smaller model can be evaluated. If the available evidence is denoted by E , the best-fitting parameter values of the saturated model are denoted by $\hat{\Theta}_K$ and those of the smaller model by $\hat{\Theta}_k$, and the probability distribution of the evidence relative to each given set of parameters is denoted by $f(E|\cdot)$, the definition of this statistic can be expressed as³

$${}_k\eta_K = -2\log \frac{f(E|\hat{\Theta}_k)}{f(E|\hat{\Theta}_K)}, \quad (1)$$

According to Bozdogan, in the limit of large samples this statistic is asymptotically distributed as a non-central χ^2 random variable with $df = K - k$ degrees of freedom and with the non-centrality parameter $N\delta$, where N is the sample size and δ is a quantity whose value does not depend on N (Bozdogan 1987, observe the change in notation). This quantity, which we shall below call the *normalized non-centrality parameter*, can be viewed as a measure of the distance between the model and the actual probability distribution of the evidence, and it has the value zero for the models which are compatible with the actual distribution (like e.g. the saturated model).

AIC is a quantity which is used for choosing between models by calculating its value for each considered model and picking the model for which this value is smallest. In the literature there are several definitions of

3. This is identical with formula (31) in Bozdogan 1987, 355 when E consists of the N independent observations x_1, x_2, \dots, x_N .

the quantity AIC, but these lead to identical choices between models. According to the most usual definition the AIC value of a model with k parameters is (see, e.g., Burnham and Anderson 1998, 46)

$$-2\log f(E|\hat{\Theta}_k) + 2k.$$

One will, of course, end up with the same model if one instead of minimizing this quantity minimizes the quantity

$$-2\log f(E|\hat{\Theta}_k) + 2k + C,$$

where C is an arbitrary constant. In particular, if the saturated model is kept fixed, the chosen model will not change if C has the value

$$C = 2\log f(E|\hat{\Theta}_K) - 2K.$$

In this case the above quantity will equal

$$\begin{aligned} AIC(k) &= -2\log f(E|\hat{\Theta}_k) + 2k + C \\ &= -2\log f(E|\hat{\Theta}_k) + 2\log f(E|\hat{\Theta}_K) + 2k - 2K \quad (2) \\ &= {}_k\eta_K - 2df \end{aligned}$$

where $df = K - k$.

I have referred to the quantity which is defined by formula (2) as AIC, because Hirotugu Akaike has suggested that in the context of factor analysis AIC should be defined to have the value which it has according to formula (2) (Akaike 1987, 321). This definition has the same contents with the one used in Mulaik 2001, 230–231.

We shall now discuss the probability distribution of $AIC(k)$. It is well-known that the non-central χ^2 random variable with df degrees of freedom and with the non-centrality parameter Δ has the expected value $\Delta + df$ and the variance $4\Delta + 2df$ (cf. Hogg and Craig 1965, 318–320). When this result is applied to the distribution of ${}_k\eta_K$, it implies that when N is large, $E({}_k\eta_K) \approx N\delta + df$ and that $D^2({}_k\eta_K) \approx 4N\delta + 2df$. Together with the definition of AIC, this implies that

$$E(AIC(k)) = E({}_k\eta_K) - 2df \approx N\delta - df,$$

and that

$$D^2(AIC(k)) = D^2({}_k\eta_K) \approx 4N\delta + 2df.$$

The result on which Mulaik bases his argument states that (Mulaik 2001, 230)

$$E(AIC(k)) \approx (N - 1)\delta - df \quad (3)$$

The reason why $N - 1$ appears here instead of N seems to be that, in a sense, the first item in the sample is worthless in the context of estimating the covariance matrix, since variances and covariances become well-defined only when the sample contains at least two items. Below I shall follow Mulaik in assuming that (3) is approximately valid, and—making a corresponding modification to the formula of the variance of $AIC(k)$ —that

$$D^2(AIC(k)) \approx 4(N - 1)\delta + 2df \quad (4)$$

The difference between these formulas and the ones which we presented before them, and which contained N in the place of $N - 1$, will be irrelevant for the discussion below. In particular, our results concerning the behavior of AIC in the limit in which $N \rightarrow \infty$ would not change if we used our earlier formulas instead of (3) and (4).

3. Comparing AIC Values in the Limit of Large Samples. Mulaik uses the approximation (3) for comparing the expected values of the AIC values of three models M_1 , M_2 , and M_3 . When AIC is used for making a choice between, e.g., the models M_1 and M_2 on the basis of a sample of some fixed size N , it will produce the methodological recommendation that the model M_1 should be preferred to the model M_2 if

$$AIC(M_1) < AIC(M_2)$$

If the symbol \mathbf{s}_N is used for denoting a sample of size N , and if the probability that the use of AIC will yield the above recommendation is denoted by $p(\{\mathbf{s}_N | AIC(M_1) < AIC(M_2)\})$, the limit of the probability that M_1 will be preferred to the model M_2 when N grows is

$$P = \lim_{N \rightarrow \infty} p(\{\mathbf{s}_N | AIC(M_1) < AIC(M_2)\}). \quad (5)$$

If this limit had the value $1/2$ for some particular models M_1 and M_2 , one could conclude AIC would not “distinguish between the two models in the limit”: in this case AIC would recommend each of the two models with almost same probability for sufficiently large samples. If, however, $P > 1/2$, AIC will recommend M_1 more often than M_2 when the sample is sufficiently large, and if $P < 1/2$, the opposite will be the case.

While discussing the question which of the two models M_1 and M_2 will be preferred in the limit, Mulaik considers their expected AIC values, and addresses the question under which circumstances it will be the case that⁴

$$E(AIC(M_1)) < E(AIC(M_2)). \quad (6)$$

It should be observed, however, that the question whether this condition is valid is quite distinct from the question how probable it is that the use of AIC leads one to choose the model M_1 , or to choose the model M_2 , when this criterion is used for choosing between the two models. By itself, the validity of the condition (6) does not imply that, if AIC is applied to the models M_1 and M_2 , it will pick more often the model M_1 than the model M_2 . This does not follow, because it is conceivable that the probability distributions of $AIC(M_1)$ and $AIC(M_2)$ were correlated in such a way that (6) was valid and $AIC(M_1)$ had nevertheless a large probability of being larger than $AIC(M_2)$. Similarly, even if it could be shown that $E(AIC(M_1)) = E(AIC(M_2))$, this would not imply that AIC would recommend the two models M_1 and M_2 equally often.

Hence, if one wants to find out which of the two models M_1 and M_2 AIC will recommend in the limit in which $N \rightarrow \infty$, one should try to find out the value of the quantity P defined by (5), rather than to find out whether the condition (6) is valid in the limit. The problem of calculating the value of P is, in general, quite difficult when M_1 and M_2 are two arbitrarily chosen factor analytic models. However, this problem becomes manageable if the model M_2 is the saturated model, as it is in the two examples which Mulaik considers.

As explained above, in the context of factor analysis the claim that M_2 is saturated implies that the estimates which M_2 yields for the numbers in the covariance matrix of the observed variables will be identical with their observed values. Denoting the normalized non-centrality parameters of the two models M_1 and M_2 by δ_1 and δ_2 , respectively, and their df values by df_1 and df_2 , respectively, it can be observed that if M_2 is saturated, $\delta_2 = df_2 = 0$. In addition, also $AIC(M_2)$ will necessarily have to be zero in this case, so that the model M_1 will be preferred to M_2 if and only if $AIC(M_1) < 0$.

Mulaik considers first a case in which the “smaller” model M_1 is, as a matter of fact, correct. This implies that also $\delta_1 = 0$, so that in this case $E(AIC(M_1)) \approx -df_1$ and $D^2(AIC(M_1)) \approx 2df_1$. If df_1 is large—which means

4. More specifically, Mulaik correctly observes that if two models M_1 and M_2 have the normalized non-centrality parameters δ_1 and δ_2 and the df values df_1 and df_2 , respectively, the approximation (3) implies that $E(AIC(M_1)) < E(AIC(M_2))$ if and only if $(N - 1)\delta_1 - df_1 < (N - 1)\delta_2 - df_2$ or, equivalently, if and only if $(\delta_1 - \delta_2) < (df_1 - df_2)/(N - 1)$. He then considers the question whether this inequality remains valid when $N \rightarrow \infty$. (See Mulaik, 2001, 230–231.)

that the number of the adjustable parameters of M_1 is essentially smaller than the number of the parameters of M_2 —it is legitimate to approximate the distribution of $AIC(M_1)$ with a normal distribution. If we denote the distribution function of the normalized normal distribution by F , we can conclude that in this case

$$\begin{aligned} P &= \lim_{N \rightarrow \infty} p(\{s_N \mid AIC(M_1) < 0\}) \approx F\left(\frac{0 - E(AIC(M_1))}{\sqrt{D^2(AIC(M_1))}}\right) \\ &= F\left(\sqrt{df_1/2}\right) \end{aligned}$$

When df_1 is large, this number will be quite close to 1, and it will be very probable that AIC yields the correct recommendation according to which M_1 should be preferred. Hence, Mulaik is mistaken when he claims that in the case that we are considering AIC would not distinguish “between a perfectly fitting model with zero δ and positive df , and a saturated model with zero δ and zero df ” (Mulaik 2001, 231).

This example is well-suited not only for illustrating the falsity of Mulaik’s claim, but also the correctness of the more moderate criticism of AIC , according to which it is not consistent. One would hope that, if the model M_1 is correct, the probability of choosing it instead of a saturated model would grow larger when the sample size increases. However, the approximate value of the probability of making the correct choice that we deduced above, $F(\sqrt{df_1/2})$, does not depend on N . This means that, even if one has collected a very large sample of observations, there is a small but positive chance of $1 - F(\sqrt{df_1/2})$ of choosing the wrong model, and this probability cannot be made to diminish by collecting still more data.

It is also natural to ask what happens when the “smaller” model M_1 is false but only “slightly off” in the sense that it is compatible with a covariance matrix which is quite close to the actual one. In this case δ_1 will have a small, positive value, and—in accordance with (3) and (4)—it will be the case that $E(AIC(M_1)) \approx (N - 1)\delta_1 - df_1$ and $D^2(AIC(M_1)) \approx 4(N - 1)\delta_1 + 2df_1$. Again, if df_1 is large, it will be legitimate to approximate the distribution of $AIC(M_1)$ with a normal distribution, and one can conclude that in this case

$$\begin{aligned} P &= \lim_{N \rightarrow \infty} p(\{s_N \mid AIC(M_1) < 0\}) \\ &= \lim_{N \rightarrow \infty} F\left(\frac{0 - [(N - 1)\delta_1 - df_1]}{\sqrt{4(N - 1)\delta_1 + 2df_1}}\right) \\ &= \lim_{N \rightarrow \infty} F\left(-\sqrt{(N - 1)\delta_1/2}\right) = 0 \end{aligned}$$

In other words, in this case the probability of choosing the smaller model which is “slightly off” will approach zero as the sample size grows, as also Mulaik correctly observes while considering this case (Mulaik 2001, 231).

This result can be motivated intuitively by considering, e.g., a case in which there are theoretical reasons for believing that the correlations between many of the observed variables are very small, and in which the model M_1 yields the result that the covariances between such observed variables is zero. In this case being “slightly off” means that there are small non-zero correlations between some of these pairs or variables. Also the fact that $P = 0$ has a simple intuitive interpretation in this case. If the empirical evidence consists of a small sample, it might be good methodology to stick to the simple model M_1 even if it is strictly speaking false. In this case the estimates of the small covariances that can be calculated using the available sample are probably worse than the estimate 0 would be. However, the situation changes as the sample size grows: if the sample is very large, the estimates of the variances and covariances which are calculated using it can be expected to be very accurate, and in this case it might be a good idea to accept these estimates as such. Hence, the choice which *AIC* in this case recommends—which is with a very great probability the saturated model, since $P = 0$ —seems to be a reasonable one in this case.

This intuitive argument would not be valid in the context of a curve-fitting problem. If a simple model—like e.g. the linear model M_{pol-1} —was false but only slightly off in the context of a curve-fitting problem, it would, of course, not be good methodology to prefer a saturated model to it. In the context of curve-fitting problems, choosing a saturated model amounts to drawing a curve which goes through all the points which represent observations, which is a rather absurd thing to do, independently of the sample size. However, the above proof of the result that that $P = 0$ when a saturated model is compared with false but “slightly off” model implicitly assumed that *the number of the parameters of the saturated model stays constant* when the size of the sample grows. This assumption is valid in the context of factor analysis, but it is not valid in the context of curve-fitting problems, because in the latter case a saturated model is a model which has the same number of parameters as there are items in the available sample. Hence, the above discussion is not applicable to curve-fitting problems, and it does not show that the use of *AIC* would be likely to lead one to prefer saturated models and their perfectly-fitting curves to the curves of models which are false, but only “slightly off.”

4. Concluding Remarks. The paper by Stanley Mulaik to which I have repeatedly referred above has broadened the scope of philosophers’ discussion concerning model selection criteria, which has otherwise been

almost exclusively concerned with curve-fitting problems, by taking an example from the field of factor analysis. My own discussion of the usefulness of *AIC* in this field has been based on the approximately valid results (3) and (4), which have been presented in the literature earlier, and in this short paper I have not made any attempt to evaluate the range within which these approximations are legitimate. Being mathematically simpler, curve-fitting examples would have suited better for illustrating the points that were made above in so far that they could have been discussed without introducing these approximations.

It would, of course, be more interesting to ask which recommendations *AIC* will yield when two ordinary, “small” models are evaluated using it, than to ask what these recommendations are when it is used for comparing a “small” model with a saturated model. When the considered models belong to the hierarchy M_{pol-0} , M_{pol-1} , M_{pol-2} , ..., the former question, which we have not addressed above, can be answered with calculations which are lengthy, but which can be carried through with elementary methods. For example, one can show—although I am for reasons of space excluding from this paper the calculations which prove that this is the case—that if the true curve belongs to the model M_{pol-0} of all horizontal straight lines, and if *AIC* is applied to choosing between M_{pol-0} and the model M_{pol-1} of all straight lines, the probability with which *AIC* will produce the recommendation that the unnecessarily large model M_{pol-1} should be used is for large samples approximately 5%. In other words, when the sample size is large and the true curve is actually a horizontal straight line, the probability with which *AIC* will correctly recommend the model which contains only horizontal straight lines is approximately 95 percent, and the probability that it will recommend the larger model which contains also all the other straight lines is approximately 5 percent.

This result is well suited for illustrating the inconsistency of *AIC* as an estimator of the number of parameters of the smallest correct model. It also illustrates the fact that the motives for introducing Akaike Information Criterion are quite different from the motives for introducing *Bayesian information criteria*. Although an easily accessible presentation of the theoretical background of *AIC* is already available (see Forster and Sober 1994), it is, perhaps, worthwhile to explain why the inconsistency of *AIC* is compatible with the results which motivate its introduction.

An argument which motivates the use of a Bayesian information criterion claims that the researcher can *maximize his probability of picking the correct model* by using the information criterion in question. Such probabilities depend on the likelihood of the evidence, given the various elements of the considered models, and on the prior probabilities of the models and their elements. However, by using *AIC* one is not supposed to maximize this probability, but something else: one is supposed to max-

imize what philosophers of science have earlier called the “*predictive accuracy*” of a chosen probability distribution, e.g. a chosen curve.

The “predictive accuracy” of a probability distribution is a quantitative measure for the distance between it and the actual distribution. In the context of a curve-fitting problem such distances are easy to visualize, since the “predictively accurate” curves are close to the correct curve. For example, if the actual distribution is represented by a horizontal line, there will be both horizontal straight lines and other curves (like straight lines which are not horizontal) which are “predictively accurate” in the sense of being very close to the correct curve. In this case the curve which actually gets chosen when *AIC* recommends a larger family of curves than M_{pol-0} will with a very great probability be very close to the correct straight line. Hence, it will with a very great probability be an acceptable choice, if the aim of the researcher is to find a curve which is “predictively accurate,” although it will be a bad choice if her aim is to find out whether the true curve is a horizontal line or not.

Above we have seen that it is incorrect to claim that *AIC* would be useless as a tool of model choice in the limit of large samples. Regarding the correct statement that *AIC* is not consistent in the limit of large samples, we can now observe that this statement is not a criticism which states that *AIC* would in some cases fail to realize to its aim. Rather, it is a criticism which states that, even if, with *AIC*, one could successfully realize the aims which are supposed to be realizable by using it, there are also other aims which a statistician could have, and one should not draw exaggerated conclusions from this success.

REFERENCES

- Akaike, Hirotugu (1987), “Factor Analysis and AIC”, *Psychometrika* 52: 317–332.
- Bollen, Kenneth A. (1989), *Structural Equations with Latent Variables*. New York: John Wiley & Sons.
- Bozdogan, Hamparsum (1987), “Model Selection and Akaike’s Information Criterion (AIC): The General Theory and its Analytic Extensions”, *Psychometrika* 52: 345–370.
- Burnham, Kenneth P., and David R. Anderson (1998), *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York: Springer.
- Forster, Malcolm, and Elliot Sober (1994), “How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions”, *British Journal for the Philosophy of Science* 45: 1–35.
- Hogg, Robert V., and Allen T. Craig (1965), *Introduction to Mathematical Statistics*. 2d ed. New York: Macmillan.
- Kiesepää, I. A. (2001), “Statistical Model Selection Criteria and the Philosophical Problem of Underdetermination”, *The British Journal for the Philosophy of Science* 52: 761–794.
- McDonald, Roderick P. (1989), “An Index of Goodness-of-Fit Based on Noncentrality”, *Journal of Classification* 6: 97–103.
- McDonald, Roderick P., and Herbert W. Marsh (1990), “Choosing a Multivariate Model: Noncentrality and Goodness of Fit”, *Psychological Bulletin* 107: 247–255.

- Mulaik, Stanley A. (2001), “The Curve-Fitting Problem: An Objectivist View”, *Philosophy of Science* 68: 218–241.
- Sakamoto, Yosiyuki, M. Ishiguro, and G. Kitagawa (1986), *Akaike Information Criterion Statistics*. Tokyo: KTK Scientific Publishers.
- Woodroffe, Michael (1982), “On Model Selection and the Arc Sine Laws”, *The Annals of Statistics* 10: 1182–1194.