

THE AKAIKE INFORMATION CRITERION AND ARTIFICIAL INTELLIGENCE

HENDRIK SCHÄBE



TÜV Rheinland InterTraffic GmbH, Germany

dr.hendrik.schaebe@gmail.com

Abstract

In this paper we use the Akaike information criterion to compare the use of parametric models with the use of very flexible models with a large parameter set, as in artificial intelligence models. It turns out that information is lost in models with many parameters, compared with using the real model.

Keywords: Akaike Information Criterion, Artificial Intelligence, Likelihood

1. INTRODUCTION

The use of artificial intelligence (AI) systems is on the rise. They are widely applied and used. So, there seems to be a big advantage of AI use. On the other hand, one must also ask: what do we lose?

In an earlier paper Schäbe and Shubinsky [1] considered what would be the correct approach to use artificial intelligence in order to gain efficiency but not to lose too much.

In this paper we will apply the Akaike information criterion to find an answer to the question, what we would lose. However, AI models are complex and there are many different classes of models. So, we can restrict ourselves to rather simple cases. AI models are mostly just statistical models, judging from the fundamental publications [2, 3], so we will use the Akaike information criterion to compare the behaviour of a AI model with the one of an ordinary statistical model. In section two, we explain the Akaike information criterion [4] and give an example, how it is applied for a kernel estimator comparing it to a parametric model. In section three, we consider a general approach. The fourth section is dedicated to the conclusions.

2. THE AKAIKE INFORMATION CRITERION

2.1. The Definition of the Akaike Information Criterion

The Akaike information criterion has been introduced by Akaike in 1974 [4]. Assume, we have a sample of size n , say $X_1, \dots, X_n$ and two different statistical models, with densities $f(x)$ and $g(x)$ having a different number of parameters k_1 and k_2 . Here θ and η are the parameters (θ being a k_1 -dimensional vector and η being a k_2 -dimensional vector) of the densities $f(x | \theta)$ and $g(x | \eta)$, whereas $\hat{\theta}$ and $\hat{\eta}$ are the coinciding values obtained as maximum likelihood estimators. Here and throughout the paper we assume that the densities exist and that they fulfil the regularity conditions as given e.g. in [5] for the maximum likelihood estimators to exist, converge in probability and being asymptotically unbiased.

Then the Akaike information criterion (AIC) compares the following quantities

$$AIC_f = k_1 - \sum_{i=1}^n \ln f(X_i | \hat{\theta}), \quad (1)$$

$$\text{AIC}_g = k_2 - \sum_{i=1}^n \ln g(X_i | \hat{\eta}), \quad (2)$$

for both models. The model with the minimal AIC value is preferred. Should the number of parameters be the same for both models, simply the model with the largest likelihood function computed for the parameter values coinciding with maximum likelihood estimators is chosen. For different numbers of parameters, a penalty term is introduced – just the number of parameters.

In the following two subsections we assume that the sample comes from a normal distribution and we will apply two different estimators. The one will be the normal distribution, i.e. the real underlying distribution of the sample. The other model is the kernel estimator. Vapnik [3] has considered this as kind of an AI application.

2.2. The AIC for a normal model

The AIC for a normal distribution model is computed as

$$\text{AIC}_N = 2k - 2 \sum_{i=1}^n \ln f(X_i | \hat{m}, \hat{\sigma}). \quad (3)$$

Since the normal distribution has 2 parameters, $k = 2$. Here $f(X_i | \hat{m}, \hat{\sigma})$ is the normal density with estimated mean and standard deviation, with

$$f(x | m, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - m)^2}{2\sigma^2}\right). \quad (4)$$

Now we compute the expectation of the AIC_N . We will neglect higher order terms in $1/n$, by partially substitution estimators by its values, e.g. for the standard deviation. Correlation between sample elements, X_i and the estimated deviation will be neglected as higher order terms. We get

$$\begin{aligned} E(\text{AIC}_N) &= 4 + 2n \ln(\sqrt{2\pi}\sigma) + nE\left[\left(\frac{X_i - m}{\sigma}\right)^2\right] \\ &= 4 + 2n \left(\ln(\sqrt{2\pi}\sigma) + \frac{1}{2}\right). \end{aligned} \quad (5)$$

The expected AIC_N for the normal distribution still depends on the variance of the distribution. Also, one can see that the AIC is proportional to n .

2.3. AIC for the Kernel Estimator

In the next step, we compute the AIC for a kernel estimator. The kernel estimator is defined by Rosenblatt and Parzen [6, 7] as

$$\tilde{f}(x) = \sum_{i=1}^n \frac{1}{nh} k\left(\frac{x - X_i}{h}\right). \quad (6)$$

For simplicity, we take a normal kernel, i.e.

$$k(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad (7)$$

which leads to

$$\tilde{f}(x) = \sum_{i=1}^n \frac{1}{nh\sqrt{2\pi}} \exp\left(-\frac{(x - X_i)^2}{2h^2}\right). \quad (8)$$

For the bandwidth we use the approximation proposed by Silverman [8], i.e.

$$h = \left(\frac{4}{3}\right)^{1/5} \hat{\sigma} n^{-1/5} \approx 1.06 \hat{\sigma} n^{-1/5}. \quad (9)$$

where $\hat{\sigma}$ is the standard deviation estimated from the sample. A more precise expression has been given by [9] using a Gram-Charlier approximation for the unknown distribution. Now, we compute the AIC_k for the kernel estimator as

$$AIC_k = 2n - 2 \sum_{i=1}^n \ln \tilde{f}(X_i). \quad (10)$$

Here the number of parameters is n , since in fact each value of the sample acts as a parameter. Since the bandwidth is derived from the standard deviation and this is also computed from the same n values X_i , the number of parameters stays n . Now we compute the expectation of AIC_k . We will again neglect higher order terms in $1/n$, by partially substitution estimators by its values, e.g. for the standard deviation. Correlation between sample elements, X_i and the estimated deviation will be neglected as higher order terms.

$$\begin{aligned} E(AIC_k) &= 2n - 2nE[\ln \tilde{f}(X_i)] \\ &= 2n + 2n \ln(nh\sqrt{2\pi}) + 2nE\left[\frac{(X_j - X_i)^2}{2h^2}\right]. \end{aligned} \quad (11)$$

$$E(AIC_k) = 2n + 2n \ln(nh\sqrt{2\pi}) + 2n \frac{\sigma^2}{h^2}. \quad (12)$$

$$\begin{aligned} E(AIC_k) &= 2n + 2n \ln(\sqrt{2\pi}) + 2n \ln\left(n^{4/5} \left(\frac{4}{3}\right)^{1/5} \sigma\right) \\ &\quad + 2n \left(\frac{4}{3n}\right)^{2/5}. \end{aligned} \quad (13)$$

$$\begin{aligned} E(AIC_k) &= 2n \left(1 + \ln(\sqrt{2\pi}) + \frac{4}{5} \ln n + \frac{1}{5} \ln \frac{4}{3} \right. \\ &\quad \left. + \ln \sigma + \left(\frac{4}{3n}\right)^{2/5}\right). \end{aligned} \quad (14)$$

$$\begin{aligned} E(AIC_k) &= 2n \left(1 + \ln(\sqrt{2\pi}) + \frac{1}{5} \ln \frac{4}{3} + \ln \sigma \right. \\ &\quad \left. + \frac{4}{5} \ln n + \left(\frac{4}{3n}\right)^{2/5}\right). \end{aligned} \quad (15)$$

$$E(AIC_k) = O(n \ln n). \quad (16)$$

2.4. Comparison

The expected AIC_k does not depend on any parameters of the underlying distribution function that we assumed to be the normal one. It can be seen that it grows with $n \ln(n)$, i.e. faster than linearly in n . Comparing this fact with the AIC for the normal model, we can see that the AIC of the kernel estimator grows faster in n than the AIC for the parametric model. The AIC shows a relative information loss and the most preferable model is always that with the smaller AIC. So we can see that with growing sample size the kernel estimator will perform worse with growing sample size, assuming the underlying model is the normal distribution.

On the one hand side this result seems to be trivial, a parametric model will always perform better than a non-parametric one – provided it is correct. On the other hand side we see that the difference in information loss for the non-parametric approach is only of the order $n^{1/5}$.

Figure 1 shows then difference of $E(AIC_k) - E(AIC_N)$. It can be seen that the difference grows with increasing n , but decreases with increasing variance of the population.

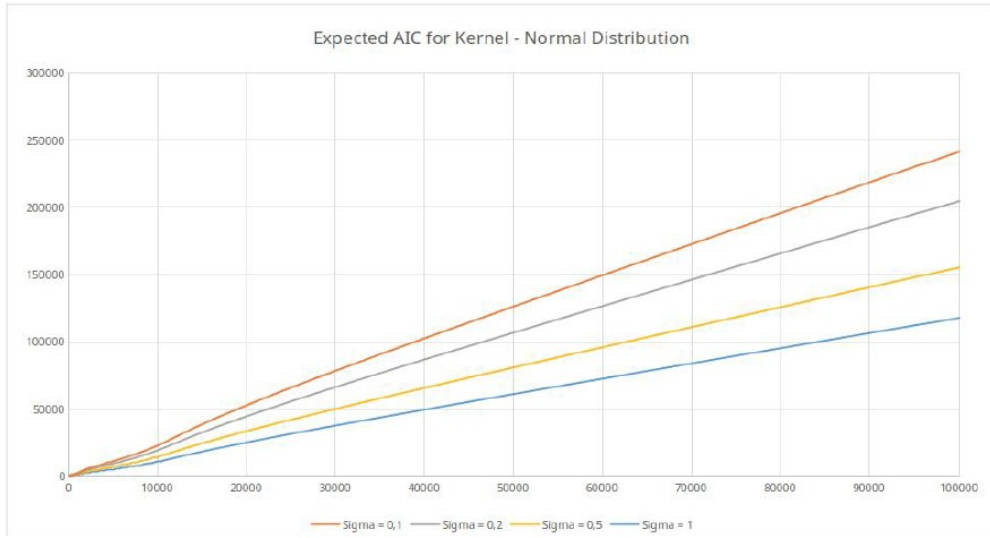


Figure 1: Comparison of the expected AIC values for the kernel estimator versus the normal model

3. A GENERALISATION

In this chapter we generalise the considerations from the previous chapter. Assume again that we have a sample $X_1, \dots, X_n$ from a distribution with density $f(x | \theta)$, where θ is a k_1 dimensional vector of parameters. The values can be multidimensional and the only restriction is, that the coinciding sigma algebra is measurable and that the density exists and fulfils the necessary regularity conditions for the existence and convergence of the maximum likelihood estimator.

Assume now that the second model is the AI model, having density $g(x | \eta)$ with parameter vector η , having k_2 dimensions and in most cases $k_2 \gg k_1$.

Now we have

$$AIC_f = k_1 - \sum_{i=1}^n \ln f(X_i | \hat{\theta}), \quad (17)$$

$$AIC_{AI} = k_2 - \sum_{i=1}^n \ln g(X_i | \hat{\eta}). \quad (18)$$

Computing the expectations we arrive at

$$E(AIC_f) = k_1 - \int_{-\infty}^{\infty} \ln f(x) f(x) dx, \quad (19)$$

$$E(AIC_{AI}) = k_2 - \int_{-\infty}^{\infty} \ln g(x) f(x) dx. \quad (20)$$

So, the difference of the expectations turns out to be

$$\begin{aligned} E(AIC_{AI}) - E(AIC_f) &= k_2 - k_1 - \int_{-\infty}^{\infty} \ln g(x) f(x) dx \\ &\quad + \int_{-\infty}^{\infty} \ln f(x) f(x) dx. \end{aligned} \quad (21)$$

The second term is just the Kullback–Leibler distance

$$D(f \parallel g) = \int_{-\infty}^{\infty} \ln \left(\frac{f(x)}{g(x)} \right) f(x) dx, \quad (22)$$

i.e.

$$E(AIC_{AI}) - E(AIC_f) = k_2 - k_1 + D(f \parallel g). \quad (23)$$

This expression is positive, if the parametric model outperforms the AI model. Now we have $k_2 > k_1$ and $D(f \parallel g) \geq 0$ for $f(x)$ and $g(x)$ being different. This shows, that using an AI model in place of the genuine parametric model leads to information loss measured by the Akaike criterion caused first by the increased number of parameters and second by using another model. This other model (here described by the density $g(x)$) might be very flexible, but is not identical to the real model $f(x)$.

4. CONCLUSION

Then result of the application of the Akaike information criterion shows, that application of an AI model always leads to information loss, compared with the use of the real correct model.

So, AI models might help for a first guess or if the real underlying model is hard to find. Nevertheless, one must always keep in mind that the use of an AI model will not use the entire information. This is also caused by the fact, as pointed out by Schäbe and Shubinsky [1], that an AI model will not use a set of sufficient statistics, which would distillate all information from the sample. The question is then, how to apply an AI model and a good first step would be to ask it, in assisting for finding the underlying model, then testing for this model and then to apply the real parametric model.

REFERENCES

- [1] Schäbe, H., and Shubinsky, I. B. (2024). How to properly apply systems of artificial intelligence. *Reliability: Theory & Applications*. 19(4):31–35. <https://doi.org/10.24412/1932-2321-2024-480-31-35>
- [2] Vapnik, V. N., and Chervonenkis, A. Y. (1974). Pattern recognition theory: Statistical problems of learning. Nauka. Moscow
- [3] Vapnik, V. N. (2010). The nature of statistical learning theory. Springer. <https://doi.org/10.1007/978-1-4757-3264-1>
- [4] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 19(6):716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- [5] Mann, N. R., et al. (1974). Methods for statistical analysis of reliability and life data. John Wiley & Sons
- [6] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*. 27(3):832–837. <https://doi.org/10.1214/aoms/1177728190>
- [7] Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*. 33(3):1065–1076. <https://doi.org/10.1214/aoms/1177704472>
- [8] Silverman, B. W. (1986). Density estimation for statistics and data analysis. Chapman & Hall/CRC. London. <https://doi.org/10.1201/9781315140919>
- [9] Schäbe, H. (1988). Zur bestimmung des glättungsparameters für kernschätzungen mit zensierten stichproben. *Wiss. Z. HfV, Dresden*. 25(3):573–591