AIC and BIC
(for model selection)

Abdelwahid Benslimane

wahid.benslimane@gmail.com

Abdelwahid Benslimane
wahid.benslimane@gmail.com

AIC

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are two indicators, similar in appearance only, which allow to compare statistical models with different numbers of parameters. The model for which these criteria have the lowest value is preferred.

The AIC and BIC are formulated as follows:

- AIC = $-2\ ln(L) +\ 2\ K$
- BIC = $-2\ ln(L) +\ K\ ln(n)$

In the above formulas, L represents the likelihood of the model to be estimated, K the number of parameters and n the number of observations.

Abdelwahid Benslimane
wahid.benslimane@gmail.com

As a reminder, the likelihood is calculated as follows (Fisher, 1920):

- $L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^{n} f(x_i|\theta)$

The best estimate of θ (parameter vector) is the one that maximizes the likelihood, i.e. the probability of having obtained the observed data.

The best model should also have a maximum likelihood, but the likelihood increases with the number of parameters, and a high number of parameters reduces capacity of generalization of a model. It is therefore necessary to find a compromise between high likelihood and low number of parameters, and it is for this purpose that the AIC and BIC are used.

Abdelwahid Benslimane
wahid.benslimane@gmail.com

The AIC approximates the Kullback-Leibler divergence between the true distribution f and the best choice in a set of parameterized models.

- $KL(f;g) = I_{fg} = \int f(x)\, ln\left(\frac{f(x)}{g(x|\theta)}\right) dx = \int f(x) ln\big(f(x)\big)dx - \int f(x) ln\big(g(x|\theta)\big)dx$

I_{fg} quantifies the information loss when g is used to approximate f.

- $I_{fg} = E_f\big[ln\big(f(x)\big)\big] - E_f\big[ln\big(g(x|\theta)\big)\big] = Constant - E_f\big[ln\big(g(x|\theta)\big)\big]$
- $I_{fg} - Constant = -E_f\big[ln\big(g(x|\theta)\big)\big]$

The quantity $-E_f\big[ln\big(g(x|\theta)\big)\big]$ is the one of interests, although it cannot be estimated, as it represents the relative directed distance between f and g. However, Hirotugu Akaike, a Japanese statistician, found that the quantity $E_\theta E_f\big[ln\big(g(x|\theta)\big)\big]$ can be estimated.

Abdelwahid Benslimane
wahid.benslimane@gmail.com

Asymptotically:

- $E_{\hat{\theta}} E_f \left[ln\left(g\left(x|\hat{\theta} \right) \right) \right] \sim ln\left(L\left(\hat{\theta} \right) \right) - K$

H. Akaike then defined the AIC in 1973 by multiplying by -2:

- AIC = $-2\ ln\left(L\left(\hat{\theta} \right) \right) + \ 2K$

When the number of parameters K is high relative to the number of observations n, i.e. if $\frac{n}{K} < 40$, it is recommended to use the corrected AIC (AIC_c).

- AIC_c = AIC + $\frac{2K(K+1)}{n-K-1}$

Abdelwahid Benslimane
wahid.benslimane@gmail.com

BIC

The BIC somehow enables Bayesian model selection.

Let us consider m models M_i of a priori distribution $P(M_i)$.

$P(\theta_i|M_i)$ is the a priori distribution of θ_i for each model. A posteriori distribution of the model given the data, $P(M_i|x)$, is proportional to the quantity $P(M_i)P(x|M_i)$.

- $P(M_i|x) \propto P(M_i)P(x|M_i) = P(M_i)\int P(x|\theta_i, M_i)P(\theta_i|M_i)d\theta_i$

With a constant prior over all m models (i.e. if all $P(M_i)$ are equal), the most probable model a posteriori is the one which maximizes $P(x|M_i)$.

Abdelwahid Benslimane
wahid.benslimane@gmail.com

After some developments we find:

- $ln\big(P(x|M_i)\big) \sim ln\left(P\left(x\middle|\hat{\theta}_i, M_i\right)\right) - \frac{K}{2} ln(n)$

Therefore, the most likely model minimizes:

- $-2\ ln\big(P(x|M_i)\big) \sim -\ 2\ ln\left(P\left(x\middle|\hat{\theta}_i, M_i\right)\right) + K\ ln(n)$ = BIC

Abdelwahid Benslimane
wahid.benslimane@gmail.com

AIC and BIC comparison

- If n tends to infinity, the probability that the BIC chooses the real model tends to 1, which is false for the AIC
- For a finite value of n, we get contradictory results. The BIC does not always choose the real model: it tends to choose models that are simple because of its stronger penalty
- The AIC will choose the model that maximizes the likelihood of future data and achieves the best bias-variance trade-off
- The AIC is a predictive criterion while the BIC is an explanatory criterion
- The BIC and AIC cannot be used simultaneously

And don’t forget: “Essentially, all models are wrong, but some are useful ” George E. P. Box