Akaike Information Criterion: The Idea[*] 2024-03-04

陳捷 Jesse C. Chen
jessekelighine.com

1 What's Wrong with Maximum Likelihood?

Suppose we have a data set $\mathbf{Y} = \{y_i\}_{i=1}^n$ and a probability density model $f(\cdot \mid \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the parameter. If we try to fit model f with the data $\mathbf{Y}$ and obtain the estimate of the parameter $\boldsymbol{\theta}$,

$$\hat{\boldsymbol{\theta}}_{\mathbf{Y}} \coloneqq \arg\max_{\boldsymbol{\theta}} \log f(\mathbf{Y} \mid \boldsymbol{\theta}). \tag{ML}$$

What are we *actually* doing here? We are supposing that *if* $\mathbf{Y}$ is generated from a probability density $f(\cdot \mid \boldsymbol{\theta}_0)$, *then* $\hat{\boldsymbol{\theta}}_{\mathbf{Y}}$ is a good estimate for $\boldsymbol{\theta}_0$. This is extensively argued by Ronald Fisher, the inventor of the Maximum Likelihood method.

Yet, this approach poses an obvious problem: *What if* $\mathbf{Y}$ follows another distribution with density function $g(\cdot \mid \boldsymbol{\phi}_0)$? We can, of course, also find the ML estimate for $\boldsymbol{\phi}_0$:

$$\hat{\boldsymbol{\phi}}_{\mathbf{Y}} \coloneqq \arg\max_{\boldsymbol{\phi}} \log g(\mathbf{Y} \mid \boldsymbol{\phi}).$$

In the spirit of ML, we can compare the two log-likelihoods,

$$\log f(\mathbf{Y} \mid \hat{\boldsymbol{\theta}}_{\mathbf{Y}}) \quad \text{and} \quad \log g(\mathbf{Y} \mid \hat{\boldsymbol{\phi}}_{\mathbf{Y}}), \tag{1}$$

and see which is larger. However, this poses another problem: since we only have one observation $\mathbf{Y}$, we can find some density function $h(\cdot \mid \boldsymbol{\psi})$ *tailored* to fit the data at hand $\mathbf{Y}$ very well, producing a high likelihood $h(\mathbf{Y} \mid \hat{\boldsymbol{\psi}}_{\mathbf{Y}})$, but fails to produce a high likelihood $h(\mathbf{X} \mid \hat{\boldsymbol{\psi}}_{\mathbf{Y}})$ when another data set $\mathbf{X}$ is presented. This is referred to as the problem of **overfitting**.

Luckily, in describing **overfitting**, we are motivated to do **cross-validation**, i.e., to use another data $\mathbf{X}$ (independent to $\mathbf{Y}$ but follows the sample distribution) to evaluate a parameter estimated under data $\mathbf{Y}$.

2 Deriving AIC

Let's switch back to using $f(\cdot \mid \boldsymbol{\theta})$ for our density function. Also let $\boldsymbol{\theta}$ be a k-dimensional vector of parameters. Instead of trying to estimate compare the log-likelihood like in (1), we try to estimate the **cross-validated** version

$$\log f(\mathbf{X} \mid \hat{\boldsymbol{\theta}}_{\mathbf{Y}}).$$

That is, after we obtained the estimator $\hat{\boldsymbol{\theta}}_{\mathbf{Y}}$ using the data set $\mathbf{Y}$, we evaluate the likelihood using another data set $\mathbf{X}$. However, since we do not have another independent data set $\mathbf{X}$, we need to do some approximation.

First, we approximate $\log f(\mathbf{X} \mid \hat{\boldsymbol{\theta}}_{\mathbf{Y}})$ by using the second-order Taylor expansion

[*] In this short introduction, I shall ignore some technical regularity conditions for clarity. I also assume the reader is familiar ML estimator, it's asymptotic properties, and Fisher information.

around $\hat{\boldsymbol{\theta}}_{\mathbf{X}}$:

$$\begin{aligned}
\log f(\mathbf{X} \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{Y}}) \approx \log f(\mathbf{X} \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{X}}) & \qquad \text{(0-th order)} \\
+ (\hat{\boldsymbol{\theta}}_{\mathbf{Y}} - \hat{\boldsymbol{\theta}}_{\mathbf{X}})^{\mathsf{T}} \left[\frac{\partial \log f(\mathbf{X} \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{X}})}{\partial \boldsymbol{\theta}} \right] & \qquad \text{(first order)} \\
+ \frac{1}{2} (\hat{\boldsymbol{\theta}}_{\mathbf{Y}} - \hat{\boldsymbol{\theta}}_{\mathbf{X}})^{\mathsf{T}} \left[\frac{\partial^2 \log f(\mathbf{X} \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{X}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} \right] (\hat{\boldsymbol{\theta}}_{\mathbf{Y}} - \hat{\boldsymbol{\theta}}_{\mathbf{X}}) & \qquad \text{(second order)}
\end{aligned}$$

Note that the first-order term (the Jacobian) is exactly zero since $\hat{\boldsymbol{\theta}}_{\mathbf{X}}$ is the ML estimator. Thus, we have

$$\begin{aligned}
\log f(\mathbf{X} \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{Y}}) \approx{} & \log f(\mathbf{X} \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{X}}) \\
& + \frac{1}{2} (\hat{\boldsymbol{\theta}}_{\mathbf{Y}} - \hat{\boldsymbol{\theta}}_{\mathbf{X}})^{\mathsf{T}} \mathbf{J}(\hat{\boldsymbol{\theta}}_{\mathbf{X}}) (\hat{\boldsymbol{\theta}}_{\mathbf{Y}} - \hat{\boldsymbol{\theta}}_{\mathbf{X}})
\end{aligned}$$

where

$$\mathbf{J}(\hat{\boldsymbol{\theta}}_{\mathbf{X}}) = \frac{\partial^2 \log f(\mathbf{X} \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{X}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}}.$$

This is the key insight of AIC: we can obtain the **cross-validated** log-likelihood by making a "correction" to the estimated likelihood $f(\mathbf{Y} \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{Y}})$. Now we split the correction term into three parts:

$$\begin{aligned}
(\hat{\boldsymbol{\theta}}_{\mathbf{Y}} - \hat{\boldsymbol{\theta}}_{\mathbf{X}})^{\mathsf{T}} \mathbf{J}(\hat{\boldsymbol{\theta}}_{\mathbf{X}}) (\hat{\boldsymbol{\theta}}_{\mathbf{Y}} - \hat{\boldsymbol{\theta}}_{\mathbf{X}}) ={} & (\hat{\boldsymbol{\theta}}_{\mathbf{X}} - \boldsymbol{\theta}_0)^{\mathsf{T}} \mathbf{J}(\hat{\boldsymbol{\theta}}_{\mathbf{X}}) (\hat{\boldsymbol{\theta}}_{\mathbf{X}} - \boldsymbol{\theta}_0) & \text{(a)} \\
& + (\hat{\boldsymbol{\theta}}_{\mathbf{Y}} - \boldsymbol{\theta}_0)^{\mathsf{T}} \mathbf{J}(\hat{\boldsymbol{\theta}}_{\mathbf{X}}) (\hat{\boldsymbol{\theta}}_{\mathbf{Y}} - \boldsymbol{\theta}_0) & \text{(b)} \\
& - 2 (\hat{\boldsymbol{\theta}}_{\mathbf{Y}} - \boldsymbol{\theta}_0)^{\mathsf{T}} \mathbf{J}(\hat{\boldsymbol{\theta}}_{\mathbf{X}}) (\hat{\boldsymbol{\theta}}_{\mathbf{X}} - \boldsymbol{\theta}_0) & \text{(c)}
\end{aligned}$$

We can easily see that part (c) goes to zero asymptotically ($n \to \infty$):

$$(\hat{\boldsymbol{\theta}}_{\mathbf{Y}} - \boldsymbol{\theta}_0)^{\mathsf{T}} \mathbf{J}(\hat{\boldsymbol{\theta}}_{\mathbf{X}}) (\hat{\boldsymbol{\theta}}_{\mathbf{X}} - \boldsymbol{\theta}_0) = \underbrace{(\hat{\boldsymbol{\theta}}_{\mathbf{Y}} - \boldsymbol{\theta}_0)^{\mathsf{T}}}_{\xrightarrow{p} 0} \mathbf{J}(\hat{\boldsymbol{\theta}}_{\mathbf{X}}) \underbrace{(\hat{\boldsymbol{\theta}}_{\mathbf{X}} - \boldsymbol{\theta}_0)}_{\xrightarrow{p} 0}.$$

Part (a) and (b) are similar in form:

$$\begin{aligned}
(\hat{\boldsymbol{\theta}}_{\mathbf{Y}} - \boldsymbol{\theta}_0)^{\mathsf{T}} \mathbf{J}(\hat{\boldsymbol{\theta}}_{\mathbf{X}}) (\hat{\boldsymbol{\theta}}_{\mathbf{Y}} - \boldsymbol{\theta}_0) &= \operatorname{trace} \Big((\hat{\boldsymbol{\theta}}_{\mathbf{Y}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_{\mathbf{Y}} - \boldsymbol{\theta}_0)^{\mathsf{T}} \mathbf{J}(\hat{\boldsymbol{\theta}}_{\mathbf{X}}) \Big) \\
(\hat{\boldsymbol{\theta}}_{\mathbf{X}} - \boldsymbol{\theta}_0)^{\mathsf{T}} \mathbf{J}(\hat{\boldsymbol{\theta}}_{\mathbf{X}}) (\hat{\boldsymbol{\theta}}_{\mathbf{X}} - \boldsymbol{\theta}_0) &= \operatorname{trace} \Big((\hat{\boldsymbol{\theta}}_{\mathbf{X}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_{\mathbf{X}} - \boldsymbol{\theta}_0)^{\mathsf{T}} \mathbf{J}(\hat{\boldsymbol{\theta}}_{\mathbf{X}}) \Big).
\end{aligned}$$

Since $\hat{\boldsymbol{\theta}}_{\mathbf{X}}$ and $\hat{\boldsymbol{\theta}}_{\mathbf{Y}}$ are both ML estimators, the blue parts above converges to the inverse of Fisher information (asymptotic variance). By information equality, $\mathbf{J}(\hat{\boldsymbol{\theta}}_{\mathbf{X}})$ also converges to the negative of Fisher information in probability. Hence, we have part (a) and (b) as the trace of identity matrices of dimension $k \times k$. That is, we have both parts equal to $-k$.

Therefore, our approximation for the **cross-validated** log-likelihood is

$$\log f(\mathbf{X} \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{Y}}) \approx \log f(\mathbf{X} \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{X}}) - k.$$

This is the famous AIC. However, AIC is often written as

$$\mathrm{AIC} = 2k - 2 \log f(\mathbf{X} \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{X}}). \qquad \text{(AIC)}$$

This is due to its connect with information theory and Kullback-Leibler Divergence.

3 AIC's Connection with Kullback-Leibler Divergence

KL divergence is an information theoretic measure of the discrepancy between two distributions. It is defined as

$$\mathrm{KL}(p \parallel q) \coloneqq \int_{\mathcal{X}} \log \left[\frac{p(x)}{q(x)}\right] p(x)\, dx$$

where p and q are two densities on the same support $\mathcal{X}$. The two main properties of KL are

1. $\mathrm{KL}(p \parallel q) \geq 0 \ \forall p, q$.
2. $\mathrm{KL}(p \parallel q) = 0$ iff $p = q$ (almost everywhere).

That is, $\mathrm{KL}(p \parallel q)$ is small when p and q are similar.

In our case, we want to know the discrepancy between the "true" likelihood function $f(\cdot \,|\, \boldsymbol{\theta}_0)$ and the estimated likelihood function $f(\cdot \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{Y}})$. Hence, we wish to choose the model with small discrepancy between the two:

$$\begin{aligned}
\mathrm{KL}(f(\cdot \,|\, \boldsymbol{\theta}_0) \parallel f(\cdot \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{Y}})) &= \int_{\mathcal{X}} \log \left[\frac{f(\mathbf{X} \,|\, \boldsymbol{\theta}_0)}{f(\mathbf{X} \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{Y}})}\right] f(\mathbf{X} \,|\, \boldsymbol{\theta}_0)\, d\mathbf{X} \\
&= \int_{\mathcal{X}} \log f(\mathbf{X} \,|\, \boldsymbol{\theta}_0) f(\mathbf{X} \,|\, \boldsymbol{\theta}_0)\, d\mathbf{X} && \text{(entropy)} \\
&+ \int_{\mathcal{X}} -\log f(\mathbf{X} \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{Y}}) f(\mathbf{X} \,|\, \boldsymbol{\theta}_0)\, d\mathbf{X} && \text{(cross-entropy)} \\
&= \text{constant} - \mathbf{E}_{\mathbf{X}} \log f(\mathbf{X} \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{Y}})
\end{aligned}$$

Thus, we can view (AIC) as an approximation of **cross-entropy**. Measuring the discrepancy between $f(\cdot \,|\, \boldsymbol{\theta}_0)$ and $f(\cdot \,|\, \hat{\boldsymbol{\theta}}_{\mathbf{Y}})$ makes intuitive sense: the problem of **overfitting** can be understood as a large discrepancy between the "true" likelihood and the "estimated" likelihood. In the original paper (Akaike, 1974), AIC is motivated by KL. Hence, AIC is represented as the *negative* of the **cross-validated** likelihood to match the sign of **cross-entropy**. Thus in practice, we want to select the model with *small* AIC.

4 Why Times Two?

If we consider a Gaussian model with $\boldsymbol{\theta} = (\mu, \sigma^2)$, the log-likelihood is written as

$$\log f(\mathbf{X} \,|\, \boldsymbol{\theta}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

It is a lot nicer to write $2 \log f(\mathbf{X} \,|\, \boldsymbol{\theta})$ so we can get rid of those $\frac{1}{2}$'s. That's why. ■

Acronyms

AIC Akaike Information Criterion. 1–3
KL Kullback-Leibler Divergence. 2, 3
ML Maximum Likelihood. 1, 2

References

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