
Akaike Information Criterion

Shuhua Hu
Center for Research in Scientific Computation
North Carolina State University
Raleigh, NC



Background

- **Model**

statistical model: $X = h(t; q) + \epsilon$

- ▼ h : mathematical model such as ODE model, PDE model, algebraic model, etc.
- ▼ ϵ : random variable with some probability distribution such as normal distribution.
- ▼ X is a random variable.

Under the assumption of ϵ being i.i.d $N(0, \sigma^2)$, we have

probability distribution model: $g(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-h(t;q))^2}{2\sigma^2}\right]$, where $\theta = (q, \sigma)$.

- ▼ g : probability density function of x depending on parameter θ .
- ▼ θ includes mathematical model parameter q and statistical model parameter σ .

- **Risk**

- ▼ **“Modeling” error (in terms of uncertainty assumption)**

Specified inappropriate parametric probability distribution for the data at hand.

▼ Estimation error

$$\|\vartheta - \hat{\theta}\|^2 = \underbrace{\|\vartheta - \theta\|^2}_{\text{bias}} + \underbrace{\|\theta - \hat{\theta}\|^2}_{\text{variance}}$$

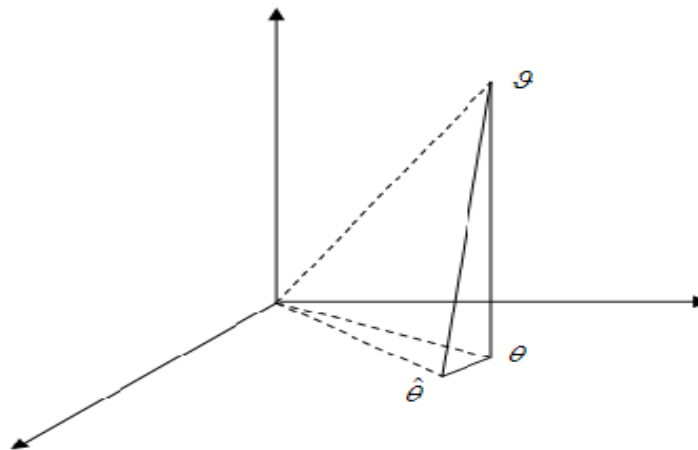
ϑ : parameter vector for the full reality model.

θ : is the projection of ϑ onto the parameter space of the approximating model Θ^k .

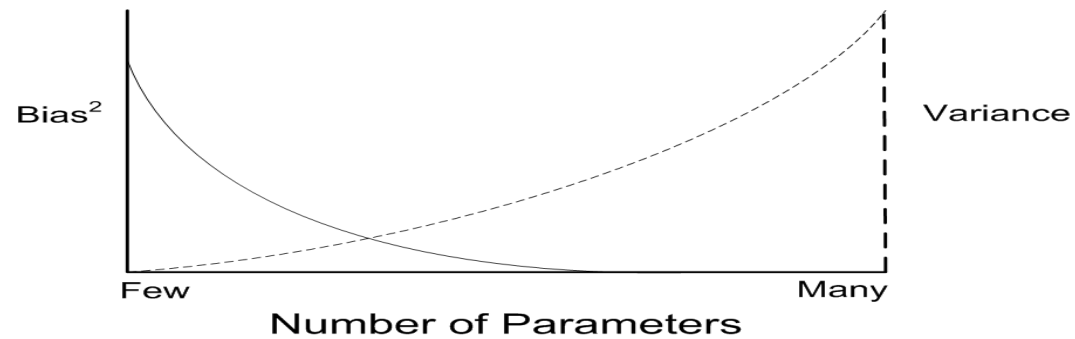
$\hat{\theta}$: the maximum likelihood estimate of θ in Θ^k .

► Variance

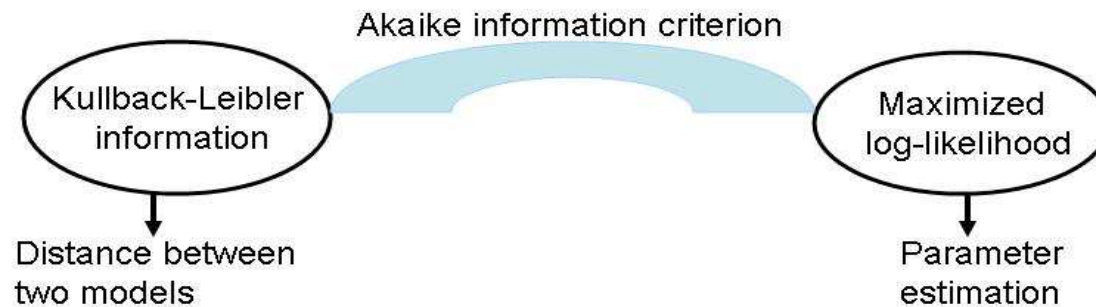
For sufficiently large sample size n , we have $n\|\theta - \hat{\theta}\|^2 \stackrel{assym.}{\sim} \chi_k^2$, where $E(\chi_k^2) = k$.



- Principle of Parsimony (*with same data set*)



- Akaike Information Criterion



Kullback-Leibler Information

Information lost when approximating model is used to approximate the full reality.

- **Continuous Case**

$$\begin{aligned}
 I(f, g(\cdot|\theta)) &= \int_{\Omega} f(x) \log \left(\frac{f(x)}{g(x|\theta)} \right) dx \\
 &= \int_{\Omega} f(x) \log(f(x)) dx - \underbrace{\int_{\Omega} f(x) \log(g(x|\theta)) dx}_{\text{relative K-L information}}
 \end{aligned}$$

▼ f : full reality or truth in terms of a probability distribution.

▼ g : approximating model in terms of a probability distribution.

▼ θ : parameter vector in the approximating model g .

- **Remark**

▼ $I(f, g) \geq 0$, with $I(f, g) = 0$ if and only if $f = g$ almost everywhere.

▼ $I(f, g) \neq I(g, f)$, which implies **K-L information is not the real “distance”**.

Akaike Information Criterion (1973)

- **Motivation**

- ▼ The truth f is unknown.

- ▼ The parameter θ in g must be estimated from the empirical data y .

- ▶ Data y is generated from $f(x)$, i.e. realization for random variable X .

- ▶ $\hat{\theta}(y)$: estimator of θ . It is a random variable.

- ▶ $I(f, g(\cdot|\hat{\theta}(y)))$ is a random variable.

- ▼ **Remark**

- ▶ We need to use expected K-L information $E_y[I(f, g(\cdot|\hat{\theta}(y)))]$ to measure the “distance” between g and f .

- **Selection Target**

Minimizing $E_y[I(f, g(\cdot|\hat{\theta}(y)))]$
 $g \in G$

$$\nabla E_y[I(f, g(\cdot|\hat{\theta}(y)))] = \int_{\Omega} f(x) \log(f(x)) dx - \underbrace{\int_{\Omega} f(y) \left[\int_{\Omega} f(x) \log(g(x|\hat{\theta}(y))) dx \right] dy}_{E_y E_x[\log(g(x|\hat{\theta}(y)))]}$$

▼ G : collection of “admissible” models (in terms of probability density functions).

▼ $\hat{\theta}$ is MLE estimate based on model g and data y .

▼ y is the random sample from the density function $f(x)$.

- **Model Selection Criterion**

Maximizing $E_y E_x[\log(g(x|\hat{\theta}(y)))]$
 $g \in G$

- **Key Result**

An approximately unbiased estimate of $E_y E_x [\log(g(x|\hat{\theta}(y)))]$ for **large sample** and **“good” model** is

$$\log(\mathcal{L}(\hat{\theta}|y)) - k$$

▼ \mathcal{L} : likelihood function.

▼ $\hat{\theta}$: maximum likelihood estimate of θ .

▼ k : number of estimated parameters (including the variance).

- **Remark**

▼ *“Good” model*: the model that is close to f in the sense of having a small K-L value.

- **Maximum Likelihood Case**

$$AIC = -2 \log \mathcal{L}(\hat{\theta}|y) + 2k$$

↙ ↘
bias **variance**

- ▼ Calculate AIC value for each model with the **same data set**, and the “best” model is the one with minimum AIC value.
- ▼ The value of AIC depends on data y , which leads to model selection uncertainty.

- **Least-Squares Case**

Assumption: i.i.d. normally distributed errors

$$AIC = n \log \left(\frac{RSS}{n} \right) + 2k$$

- ▼ RSS is estimated residual of fitted model.

Takeuchi's Information Criterion (1976)

useful in cases where the model is not particular close to truth.

- **Model Selection Criterion**

$$\text{Maximizing } E_y E_x [\log(g(x|\hat{\theta}(y)))]_{g \in G}$$

- **Key Result**

An approximately unbiased estimator of $E_y E_x [\log(g(x|\hat{\theta}(y)))]$ for **large sample** is

$$\log(\mathcal{L}(\hat{\theta}|y)) - \text{tr}(J(\theta_0)I(\theta_0)^{-1})$$

$$\blacktriangledown J(\theta_0) = E_f \left[\left(\frac{\partial}{\partial \theta} \log(g(x|\theta)) \right) \left(\frac{\partial}{\partial \theta} \log(g(x|\theta)) \right)^T \right]_{|\theta=\theta_0}$$

$$\blacktriangledown I(\theta_0) = E_f \left[-\frac{\partial^2 \log(g(x|\theta))}{\partial \theta_i \partial \theta_j} \right]_{|\theta=\theta_0}$$

- **Remark**

▼ If $g \equiv f$, then $I(\theta_0) = J(\theta_0)$. Hence $\text{tr}(J(\theta_0)I(\theta_0)^{-1}) = k$.

▼ If g is close to f , then $\text{tr}(J(\theta_0)I(\theta_0)^{-1}) \approx k$.

- **TIC**

$$TIC = -2 \log(\mathcal{L}(\hat{\theta}|y)) + 2\text{tr}(\hat{J}(\hat{\theta})[\hat{I}(\hat{\theta})]^{-1}),$$

where $\hat{I}(\hat{\theta})$ and $\hat{J}(\hat{\theta})$ are both $k \times k$ matrix, and

$$\hat{I}(\hat{\theta}) = -\frac{\partial^2 \log(g(x|\hat{\theta}))}{\partial \theta^2} \rightarrow \text{estimate of } I(\theta_0)$$

$$\hat{J}(\hat{\theta}) = \sum_{i=1}^n \left[\frac{\partial}{\partial \theta} \log(g(x_i|\hat{\theta})) \right] \left[\frac{\partial}{\partial \theta} \log(g(x_i|\hat{\theta})) \right]^T \rightarrow \text{estimate of } J(\theta_0)$$

- **Remark**

- ▼ Attractive in theory.

- ▼ Rarely used in practice because we need a very large sample size to obtain good estimates for both $I(\theta_0)$ and $J(\theta_0)$.

A Small Sample AIC

use in the case where the sample size is small relative to the number of parameters
rule of thumb: $n/k < 40$

- **Univariate Case**

Assumption: i.i.d normal error distribution with the truth contained in the model set.

$$AIC_c = AIC + \underbrace{\frac{2k(k+1)}{n-k-1}}_{\text{bias-correction}}$$

- **Remark**

- ▼ The bias-correction term varies by type of model (e.g., normal, exponential, Poisson).
- ▼ In practice, AIC_c is generally suitable unless the underlying probability distribution is extremely nonnormal, especially in terms of being strongly skewed.

- **Multivariate Case**

Assumption: each row of ϵ is i.i.d $\mathbf{N}(0, \Sigma)$.

$$AIC_c = AIC + 2 \frac{k(\tilde{k} + 1 + p)}{n - \tilde{k} - 1 - p}$$

▼ Applying to the multivariate case:

$$Y = TB + \epsilon, \text{ where } Y \in \mathbb{R}^{n \times p}, T \in \mathbb{R}^{n \times \tilde{k}}, B \in \mathbb{R}^{\tilde{k} \times p}.$$

▼ p : total number of components.

▼ n : number of independent multivariate observations, each with p nonindependent components.

▼ k : total number of unknown parameters and $k = \tilde{k}p + p(p + 1)/2$.

- **Remark**

▼ Bedrick and Tsai in [1] claimed that this result can be extended to the multivariate non-linear regression model.

AIC Differences, Likelihood of a Model, Akaike Weights

- **AIC differences**

Information loss when fitted model is used rather than the best approximating model

$$\Delta_i = AIC_i - AIC_{\min}$$

▼ AIC_{\min} : AIC values for the best model in the set.

- **Likelihood of a Model**

Useful in making inference concerning the relative strength of evidence for each of the models in the set

$$\mathcal{L}(g_i|y) \propto \exp\left(-\frac{1}{2}\Delta_i\right), \text{ where } \propto \text{ means "is proportional to".}$$

- **Akaike Weights**

"Weight of evidence" in favor of model i being the best approximating model in the set

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^R \exp(-\frac{1}{2}\Delta_r)}$$

Confidence Set for K-L Best Model

- **Three Heuristic Approaches** (see [4])

- ▼ **Based on the Akaike weights w_i**

To obtain a 95% confidence set on the actual K-L best model, summing the Akaike weights from largest to smallest until that sum is just ≥ 0.95 , and the corresponding subset of models is the confidence set on the K-L best model.

- ▼ **Based on AIC difference Δ_i**

- ▶ $0 \leq \Delta_i \leq 2$, substantial support,
- ▶ $4 \leq \Delta_i \leq 7$, considerable less support,
- ▶ $\Delta_i > 10$, essentially no support.

Remark

- ▶ Particularly useful for nested models, may break down when the model set is large.
- ▶ The guideline values may be somewhat larger for nonnested models.

- ▼ **Motivated by likelihood-based inference**

The confidence set of models is all models for which the ratio

$$\frac{\mathcal{L}(g_i|y)}{\mathcal{L}(g_{\min}|y)} > \alpha, \text{ where } \alpha \text{ might be chosen as } \frac{1}{8}.$$

Multimodel Inference

- **Unconditional Variance Estimator**

$$\widehat{\text{var}}(\hat{\theta}) = \left[\sum_{i=1}^R w_i \sqrt{\widehat{\text{var}}(\hat{\theta}_i | g_i) + (\hat{\theta}_i - \hat{\theta})^2} \right]^2$$

- ▼ θ is a parameter in common to all R models.
- ▼ $\hat{\theta}_i$ means that the parameter θ is estimated based on model g_i ,
- ▼ $\hat{\theta}$ is a model-averaged estimate $\hat{\theta} = \sum_{i=1}^R w_i \hat{\theta}_i$.

- **Remark**

- ▼ “Unconditional” means not conditional on any particular model, but still conditional on the full set of models considered.
- ▼ If θ is a parameter in common to only a subset of the R models, then w_i must be recalculated based on just these models (thus these new weights must satisfy $\sum w_i = 1$).
- ▼ Use unconditional variance unless the selected model is strongly supported (for example, $w_{\min} > 0.9$).

Summary of Akaike Information Criteria

- **Advantages**

- ▼ Valid for both nested and nonnested models.
- ▼ Compare models with different error distribution.
- ▼ Avoid multiple testing issues.

- **Selected Model**

- ▼ The model with minimum AIC value.
- ▼ Specific to given data set.

- **Pitfall in Using Akaike Information Criteria**

- ▼ **Can not be used to compare models of different data sets.**

For example, if nonlinear regression model g_1 is fitted to a data set with $n = 140$ observations, one cannot validly compare it with model g_2 when 7 outliers have been deleted, leaving only $n = 133$.

▼ **Should use the same response variables for all the candidate models.**

For example, if there was interest in the normal and log-normal model forms, the models would have to be expressed, respectively, as

$$g_1(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right], \quad g_2(x|\mu, \sigma) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left[-\frac{(\log(x) - \mu)^2}{2\sigma^2}\right],$$

instead of

$$g_1(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right], \quad g_2(\log(x)|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\log(x) - \mu)^2}{2\sigma^2}\right].$$

▼ **Do not mix null hypothesis testing with information criterion.**

- ▶ Information criterion is not a “test”, so avoid use of “significant” and “not significant”, or “rejected” and “not rejected” in reporting results.
- ▶ Do not use *AIC* to rank models in the set and then test whether the best model is “significantly better” than the second-best model.

▼ **Should retain all the components of each likelihood in comparing different probability distributions.**

References

- [1] E.J. Bedrick and C.L. Tsai, Model Selection for Multivariate Regression in Small Samples, *Biometrics*, 50 (1994), 226–231.
- [2] H. Bozdogan, Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions, *Psychometrika*, 52 (1987), 345–370.
- [3] H. Bozdogan, Akaike's Information Criterion and Recent Developments in Information Complexity, *Journal of Mathematical Psychology*, 44 (2000), 62–91.
- [4] K.P. Burnham and D.R. Anderson, *Model Selection and Inference: A Practical Information-Theoretical Approach*, (1998), New York: Springer-Verlag.
- [5] K.P. Burnham and D.R. Anderson, Multimodel Inference: Understanding AIC and BIC in Model Selection, *Sociological methods and research*, 33 (2004), 261–304.
- [6] C.M. Hurvich and C.L. Tsai, Regression and Time Series Model Selection in Small Samples, *Biometrika*, 76 (1989).