Information criteria

how to build and interpret them

Model quality

- Consider
  - true distribution g(x)
  - model distribution f(x)
- Popular model quality estimates:
  - Kullback-Leibler divergence
  - generalized information
  - L^2 norm
  - *etc.*

KL divergence

$$KL(g(x)\|f(x)) := \int g(x)\log\left(\frac{g(x)}{f(x)}\right)dx =$$

$$= E_{X\sim g(x)}\left[\log\left(\frac{g(X)}{f(X)}\right)\right] =$$

$$= E_{X\sim g(x)}[\log g(X) - \log f(X)]$$

- Also called *KL information* and (incorrectly) *KL distance*
- KL = 0 **iff** g(x) = f(x); KL > 0 **otherwise**
  - standard information properties

Model selection

- Model with minimal KL is better.
- Direct computation of KL is impossible:
  - we don't know g(x)
- Use various approximations.
- Suffices to compare the following value:

$$-E_{X\sim g(x)}[\log f(X)]$$

Fixed model selection

Approximate the KL-summand using empirical distribution given by the sample set:

$$-E_{X\sim g(x)}[\log f(X)]\approx -E_{X\sim \hat{g}(x)}[\log f(X)]=$$

$$=-\frac{1}{n}\sum_{k}\log f(X_k)=-\frac{1}{n}\boldsymbol{L}(\vec{X})$$

Fixed model selection

- Good news:
  the average sample log-likelihood is a good estimator for ***the expected log-likelihood***
  - use central limit theorem to establish estimator consistence
- Use the sample log-likelihood to select the best model from the fixed model set.

MLE models

- This approach doesn't work for MLE models:
  - the model distribution f(x) is sample-dependent
  - thus the estimator is biased
- Correction for bias is needed:

$$\delta(\vec{X}) = \log f(\vec{X}|\hat{\theta}(\vec{X})) - n E_{Z\sim g(x)}[\log f(Z|\hat{\theta}(\vec{X}))]$$

$$b = E_{\vec{X}\sim g(\vec{x})}\delta(\vec{X})$$

IC for MLE models

The general form of an information criterion for MLE model selection is thus:

$$IC(\vec{X}, f) = -2\log f(\vec{X}|\hat{\theta}(\vec{X})) + 2\hat{b}(\vec{X}, f)$$

where

$$\hat{b} \approx b = E_{\vec{X} \sim g(\vec{x})}[\delta(\vec{X})]$$

is an estimator for the bias.

Bias approximation

Let

$$\theta_0 := \arg\max_\theta E_{X\sim g(x)}[\log f(X|\theta)]$$

then

$$b \approx \mathrm{tr}[I(\theta_0)J(\theta_0)^{-1}]$$

where I is *Fisher information,*
and J is *the negative Hessian matrix of* $\log f(X|\theta)$.

This approximation works asymptotically,
provided continuous differentiability.

TIC

We still can't compute θ_0, so we need to estimate this value too.

$\hat{\theta}=\hat{\theta}(\vec{X})$ seems a natural choice.

We also need to replace expectation in the Fisher and Hessian with sample average.

Takeuchi information criterion:

$$TIC(\vec{X},f)=-2\log f(\vec{X}|\hat{\theta}(\vec{X}))+2\,\mathrm{tr}[\hat{I}(\hat{\theta})\hat{J}^{-1}(\hat{\theta})]$$

AIC

Suppose that

$$g(x) = f(x|\theta_0)$$

for some θ_0. Then

$$I(\theta_0) = J(\theta_0)$$

and

$$b \approx \mathrm{tr}\left[I(\theta_0)J(\theta_0)^{-1}\right] = \mathrm{tr}\,\mathrm{E} = \dim\theta$$

AIC

Akaike information criterion:

$$AIC = -2 \log f\left(\vec{X} | \hat{\theta}\left(\vec{X}\right)\right) + 2 \dim \theta$$

And now for something completely different

- Consider Bayesian formalism
- We need to determine:

$$P\left(\vec{X}|\text{model}\right)=\int_{\theta} P\left(\vec{X}|\theta,\text{model}\right)P\left(\theta|\text{model}\right)d\,\theta$$

- We can determine:

$$\max_{\theta} P\left(\vec{X}|\theta,\text{model}\right)$$

Bayesian formalism

- Integration is hard (and usually intractable)
- Various numeric techniques exist
- However, one can use Laplace's method for approximation
- Approximation prerequisites:
  - single global distribution maximum
  - thin tail (?)
  - flat parameter prior

Laplace's trick

Laplace's method:

$$\int_a^b \exp\left(n f(x)\right) dx \underset{n\to\infty}{\simeq} \exp\left(n f(x_0)\right) \sqrt{\frac{2\pi}{n\left|f''(x_0)\right|}}$$

where x_0 is the global maximum point.

Laplace's trick

Laplace's method for multivariate functions:

$$\int_U \exp\left(n f\left(\vec{x}\right)\right) d\vec{x} \underset{n\to\infty}{\sim} \exp\left(n f\left(\vec{x}_0\right)\right)\left(\frac{2\pi}{n\left|-H_f\left(\vec{x}_0\right)\right|}\right)^{\frac{k}{2}}$$

Here $k = \dim \vec{x}$ and H_f is the Hessian matrix.

Laplace's trick

In our case:

$$P(\vec{X}) = \int \exp(\log P(\vec{X}|\theta)) P(\theta) d\theta$$
$$= C \int \exp\left(n \cdot \frac{1}{n} \log P(\vec{X}|\theta) \right) d\theta$$

Note that:

$$\frac{1}{n} \log P(\vec{X}|\theta) = \frac{\sum \log P(X_k|\theta)}{n} \rightarrow \varphi(\theta)$$

Laplace's trick

The MLE $\hat{\theta}=\hat{\theta}\left(\vec{X}\right)$ is usually a consistent estimator for $\theta_0=\arg\max\varphi\left(\theta\right)$.

Thus

$$P\left(\vec{X}\right)\sim C\exp\left(n\varphi\left(\hat{\theta}\right)\right)\left(\frac{2\pi}{n\left|-H_{\varphi}\left(\theta_0\right)\right|}\right)^{\frac{k}{2}}$$

It's worth noting that $-H_{\varphi}\left(\theta_0\right)=I\left(\theta_0\right)$, where I is the Fisher information matrix.

BIC

- In this case the following asymptotic holds:

$$\log P\left(\vec{X}\right) \sim \max_{\theta} \log P\left(\vec{X}|\theta\right) - \frac{k}{2} \cdot \log n$$

- Bayesian information criterion:

$$BIC = k \cdot \log n - 2 \max_{\theta} \log P\left(\vec{X}|\theta\right)$$

Conclusion (sort of)

- AIC and BIC are mostly silver bullets used when the distribution is too complex (or the researcher is too lazy)
- However, they are good for prototypic research:
  - building more precise criteria takes some time,
  - using standard IC can discard the hypothesis before they are needed.