Model Selection Tutorial #1: Akaike’s Information Criterion

Daniel F. Schmidt and Enes Makalic

Melbourne, November 22, 2008
1 Motivation

2 Estimation

3 AIC

4 Derivation

5 References
We have observed $n$ data points $y^n = (y_1, \ldots, y_n)$ from some unknown, probabilistic source $p^*$, i.e.

$$y^n \sim p^*$$

where $y^n \in \mathcal{Y}^n$.

We wish to learn about $p^*$ from $y^n$.

More precisely, we would like to discover the generating source $p^*$, or at least a good approximation of it, from nothing but $y^n$. 
Statistical Models

- To approximate $p^*$ we will restrict ourselves to a set of potential statistical models.
- Informally, a statistical model can be viewed as a conditional probability distribution over the potential dataspace $\mathcal{Y}^n$

$$p(y^n|\theta), \ \theta \in \Theta$$

where $\theta = (\theta_1, \ldots, \theta_p)$ is a parameter vector that indexes the particular model.

- Such models satisfy

$$\int_{y^n \in \mathcal{Y}^n} p(y^n|\theta)dy^n = 1$$

for a fixed $\theta$. 

Daniel F. Schmidt and Enes Makalic  
Model Selection with AIC
An example would be the univariate normal distribution.

\[ p(y^n|\theta) = \left( \frac{1}{2\pi\tau} \right)^{\frac{n}{2}} \exp \left( -\frac{1}{2\tau} \sum_{i=1}^{n} (y_i - \mu)^2 \right) \]

where

- \( p = 2 \)
- \( \theta = (\mu, \tau) \) are the parameters
- \( Y^n = \mathbb{R}^n \)
- \( \Theta = \mathbb{R} \times \mathbb{R}_+ \)
Motivation

Estimation

AIC

Derivation

References

Daniel F. Schmidt and Enes Makalic

Model Selection with AIC
Given a statistical model and data $y^n$, we would like to take a guess at a plausible value of $\theta$.

The guess should be ‘good’ in some sense.

Many ways to approach this problem; we shall discuss one particularly relevant and important method: Maximum Likelihood.
A heuristic procedure introduced by R. A. Fisher
Possesses good properties in many cases
Is very general and easy to understand
To estimate parameters $\theta$ for a statistical model from $y^n$, solve
\[
\hat{\theta}(y^n) = \arg \max_{\theta \in \Theta} \{ p(y^n | \theta) \}
\]
or, more conveniently
\[
\hat{\theta}(y^n) = \arg \min_{\theta \in \Theta} \{ -\log p(y^n | \theta) \}
\]
Example: Estimating the mean parameter $\mu$ of a univariate normal distribution

Negative log-likelihood function:

$$L(\mu, \tau) = \frac{n}{2} \log(2\pi \tau) + \frac{1}{2\tau} \sum_{i=1}^{n} (y_i - \mu)^2$$
Example: Estimating the mean parameter $\mu$ of a univariate normal distribution.

Negative log-likelihood function:

$$L(\mu, \tau) = \frac{n}{2} \log(2\pi\tau) + \frac{1}{2\tau} \sum_{i=1}^{n} (y_i - \mu)^2$$

Differentiating $L(\cdot)$ with respect to $\mu$ yields

$$\frac{\partial L(\mu, \tau)}{\partial \mu} = \frac{1}{2\tau} \left( 2n\mu - 2 \sum_{i=1}^{n} y_i \right)$$
Example: Estimating the mean parameter $\mu$ of a univariate normal distribution

Negative log-likelihood function:

$$L(\mu, \tau) = \frac{n}{2} \log(2\pi\tau) + \frac{1}{2\tau} \sum_{i=1}^{n} (y_i - \mu)^2$$

Differentiating $L(\cdot)$ with respect to $\mu$ yields

$$\frac{\partial L(\mu, \tau)}{\partial \mu} = \frac{1}{2\tau} \left( 2n\mu - 2 \sum_{i=1}^{n} y_i \right)$$

Setting this to zero, and solving for $\mu$ yields

$$\hat{\mu}(y^n) = \frac{1}{n} \sum_{i=1}^{n} y_i$$
A more complex model: $k$-order polynomial regression
Univariate Polynomial Regression

- A more complex model: $k$-order polynomial regression
- Let each $y(x)$ be distributed as per a univariate normal with variance $\tau$ and a special mean

\[ \mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \ldots \beta_k x^k \]

The parameters of this model are $\theta^{(k)} = (\tau, \beta_0, \ldots, \beta_k)$. 
A more complex model: \( k \)-order polynomial regression

Let each \( y(x) \) be distributed as per a univariate normal with variance \( \tau \) and a special mean

\[
\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \ldots \beta_k x^k
\]

The parameters of this model are \( \theta^{(k)} = (\tau, \beta_0, \ldots, \beta_k) \).

In this model the data \( y^n \) is associated with a \( x^n \) which are known.
A more complex model: $k$-order polynomial regression

Let each $y(x)$ be distributed as per a univariate normal with variance $\tau$ and a special mean

$$\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \ldots \beta_k x^k$$

The parameters of this model are $\theta^{(k)} = (\tau, \beta_0, \ldots, \beta_k)$.

In this model the data $y^n$ is associated with a $x^n$ which are known.

Given an order $k$, maximum likelihood can be used to estimate $\theta^{(k)}$.
A more complex model: $k$-order polynomial regression

Let each $y(x)$ be distributed as per a univariate normal with variance $\tau$ and a special mean

$$\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \ldots \ldots \beta_k x^k$$

The parameters of this model are $\theta^{(k)} = (\tau, \beta_0, \ldots, \beta_k)$.

In this model the data $y^n$ is associated with a $x^n$ which are known

**Given** an order $k$, maximum likelihood can be used to estimate $\theta^{(k)}$

But it cannot be used to provide a suitable estimate of order $k$!
Univariate Polynomial Regression

If we let

\[ \hat{\mu}^{(k)}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 \ldots \ldots \hat{\beta}_k x^k \]

Maximum Likelihood chooses \( \hat{\beta}^{(k)}(y^n) \) to minimise

\[ \hat{\tau}^{(k)}(y^n) = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \hat{\mu}^{(k)}(x_i) \right)^2 \]

This is called the \textit{residual variance}. 
Univariate Polynomial Regression

- If we let

\[ \hat{\mu}^{(k)}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 \ldots \hat{\beta}_k x^k \]

Maximum Likelihood chooses \( \hat{\beta}^{(k)}(y^n) \) to minimise

\[ \hat{\tau}^{(k)}(y^n) = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \hat{\mu}^{(k)}(x_i) \right)^2 \]

This is called the residual variance.

- The likelihood function \( L(y^n|\hat{\theta}^{(k)}(y^n)) \) made by plugging in the Maximum Likelihood estimates is

\[ L(y^n|\hat{\theta}^{(k)}(y^n)) = \frac{n}{2} \log \left( 2\pi \hat{\tau}^{(k)}(y^n) \right) + \frac{n}{2} \]
‘Truth’ : $\mu(x) = 9.7x^5 + 0.8x^3 + 9.4x^2 - 5.7x - 2$, $\tau = 1$
Polynomial fit, $k = 2$, $\hat{f}^{(2)}(y) = 4.6919$
Polynomial fit, $k = 5$, $\hat{\tau}^{(5)}(y) = 1.1388$
Polynomial fit, $k = 10$, $\hat{f}^{(10)}(y) = 1.0038$
Polynomial fit, $k = 20$, $\hat{f}^{(20)}(y) = 0.1612$
A problem with Maximum Likelihood

It is not difficult to show that

\[ \hat{\tau}^{(0)} > \hat{\tau}^{(1)} > \hat{\tau}^{(2)} > \ldots > \hat{\tau}^{(n-1)} \]

and furthermore that \( \hat{\tau}^{(n-1)} = 0. \)
It is not difficult to show that

$$\hat{\tau}^{(0)} > \hat{\tau}^{(1)} > \hat{\tau}^{(2)} > \ldots > \hat{\tau}^{(n-1)}$$

and furthermore that $\hat{\tau}^{(n-1)} = 0$.

From this it is obvious that attempting to estimate $k$ using Maximum Likelihood will fail, i.e. the solution of

$$\hat{k} = \arg \min_{k \in \{0, \ldots, n-1\}} \left\{ \frac{n}{2} \log 2\pi \hat{\tau}^{(k)}(y^n) + \frac{n}{2} \right\}$$

is simply $\hat{k} = (n - 1)$, irrespective of $y^n$. 
Some solutions ...

- The minimum encoding approach, pioneered by C.S. Wallace, D. Boulton and J.J. Rissanen

- The minimum discrepancy estimation approach, pioneered by H. Akaike
AIC is based on estimating the Kullback-Leibler (KL) divergence.

The Kullback-Leibler divergence

\[
KL(f \| g) = - \int_{\gamma^n} f(y^n) \log g(y^n) dy^n + \int_{\gamma^n} f(y^n) \log f(y^n) dy^n
\]

- Cross-entropy,
- Entropy

Cross-entropy, \( \Delta(f \| g) \), is the ‘expected negative log-likelihood’ of data coming from \( f \) under \( g \).
Cross-entropy for polynomial fits of order $k = \{0, \ldots, 20\}$
Akaike’s Information Criterion

- Problem: KL divergence depends on knowing the truth (our $p^*$)

- Akaike’s solution: Estimate it!
The AIC score for a model is

$$AIC(\hat{\theta}(y^n)) = - \log p(y^n|\hat{\theta}(y^n)) + p$$

where $p$ is the number of free model parameters.

Using AIC one chooses the model that solves

$$\hat{k} = \arg \min_{k \in \{0,1,...\}} \left\{AIC(\hat{\theta}^{(k)}(y^n))\right\}$$
Under certain conditions the AIC score satisfies

\[ E_{\theta^*} \left[ \text{AIC}(\hat{\theta}(y^n)) \right] = E_{\theta^*} \left[ \Delta(\theta^* \mid \mid \hat{\theta}(y^n)) \right] + o_n(1) \]

where \( o_n(1) \to 0 \) as \( n \to \infty \)

In words, the AIC score is an \textit{asymptotically unbiased} estimate of the cross-entropy risk

This means it is only valid if \( n \) is ‘large’
Properties of AIC

- AIC is good for prediction
- AIC is an *asymptotically efficient* model selection criterion
- In words, as $n \to \infty$, with probability approaching one, the model with the minimum AIC score will also possess the smallest Kullback-Leibler divergence
- It is not necessarily the best choice for *induction*
AIC is an asymptotic approximation; one should consider whether it applies before using it.
Conditions for AIC to apply

- AIC is an asymptotic approximation; one should consider whether it applies before using it.
- For AIC to be valid, $n$ must be large compared to $p$. 

Conditions for AIC to apply

- AIC is an asymptotic approximation; one should consider whether it applies before using it.
- For AIC to be valid, \( n \) must be large compared to \( p \).
- The true model must be \( \theta^* \in \Theta \).
AIC is an asymptotic approximation; one should consider whether it applies before using it.

For AIC to be valid, $n$ must be large compared to $p$.

The true model must be $\theta^* \in \Theta$.

Every $\theta \in \Theta$ must map to a unique distribution $p(\cdot | \theta)$.
AIC is an asymptotic approximation; one should consider whether it applies before using it.

For AIC to be valid, $n$ must be large compared to $p$.

The true model must be $\theta^* \in \Theta$.

Every $\theta \in \Theta$ must map to a unique distribution $p(\cdot | \theta)$.

The Maximum Likelihood estimates must be consistent and be approximately normally distributed for large $n$. 
AIC is an asymptotic approximation; one should consider whether it applies before using it.

For AIC to be valid, \( n \) must be large compared to \( p \).

The true model must be \( \theta^* \in \Theta \).

Every \( \theta \in \Theta \) must map to a unique distribution \( p(\cdot|\theta) \).

The Maximum Likelihood estimates must be consistent and be approximately normally distributed for large \( n \).

\( L(\theta) \) must be twice differentiable with respect to \( \theta \) for all \( \theta \in \Theta \).
Some models to which AIC can be applied include...

- Linear regression models, function approximation
- Generalised linear models
- Autoregressive Moving Average models, spectral estimation
- Constant bin-width histogram estimation
- Some forms of hypothesis testing
When not to use AIC

- Multilayer Perceptron Neural Networks
  - Many different $\theta$ map to the same distribution
When not to use AIC

- Multilayer Perceptron Neural Networks
  - Many different $\theta$ map to the same distribution
- Neyman-Scott Problem, Mixture Modelling
  - The Maximum Likelihood estimates are not consistent
When not to use AIC

- **Multilayer Perceptron Neural Networks**
  - Many different $\theta$ map to the same distribution

- **Neyman-Scott Problem, Mixture Modelling**
  - The Maximum Likelihood estimates are not consistent

- **The Uniform Distribution**
  - $L(\theta)$ is not twice differentiable
When not to use AIC

- Multilayer Perceptron Neural Networks
  - Many different $\theta$ map to the same distribution
- Neyman-Scott Problem, Mixture Modelling
  - The Maximum Likelihood estimates are not consistent
- The Uniform Distribution
  - $L(\theta)$ is not twice differentiable
- The AIC approach may still be applied to these problems, but the derivations need to be different
Application to polynomials

- **AIC criterion for polynomials**

\[
AIC(k) = \frac{n}{2} \log 2\pi \hat{\tau}(k) (\mathbf{y}^n) + \frac{n}{2} + (k + 2)
\]
AIC selects $\hat{k} = 3$
For some model types it is possible to derive improved estimates of the cross-entropy

Under certain conditions, the ‘corrected’ AIC (AICc) criterion

\[
AIC_c(\hat{\theta}(y^n)) = -\log p(y^n|\hat{\theta}(y^n)) + \frac{n(p + 1)}{n - p - 2}
\]

satisfies

\[
E_{\theta^*} \left[ AIC_c(\hat{\theta}(y^n)) \right] = E_{\theta^*} \left[ \Delta(\theta^* \parallel \hat{\theta}(y^n)) \right]
\]

In words, it is an exactly unbiased estimator of the cross-entropy, even for finite \( n \)
Application to polynomials

- AICc criterion for polynomials

\[
AIC_c(k) = \frac{n}{2} \log 2\pi \hat{\tau}^{(k)}(y^n) + \frac{n}{2} + \frac{n(k + 2)}{n - k - 3}
\]
Using AICc

- Tends to perform better than AIC, especially when $n/p$ is small
- Theoretically only valid for homoskedastic linear models; these include
  - Linear regression models, including linear function approximation
  - Autoregressive Moving Average (ARMA) models
  - Linear smoothers (kernel, local regression, etc)
- Practically, tends to perform well as long as the model class is suitably regular
Some theory

- Let $k^*$ be the true number of parameters, and assume that the model space is nested.
- Two sources of error/discrepancy in model selection.
  - Discrepancy due to approximation:
    - Main source of error when underfitting, i.e. when $\hat{k} < k^*$.
  - Discrepancy due to estimation:
    - Source of error when exactly fitting or overfitting, i.e. when $\hat{k} \geq k^*$.
Discrepancy due to Approximation

![Graph showing the discrepancy between the true curve and the best fitting cubic curve.](image-url)
Discrepancy due to Estimation

![Graph showing discrepancies due to estimation.](image)
The aim is to show that

\[ E_{\theta^*} \left[ L(y^n|\hat{\theta}) + \rho \right] = E_{\theta^*} \left[ \Delta(\theta^*||\hat{\theta}) \right] + o_n(1) \]
The aim is to show that

$$E_{\theta^*} \left[ L(y^n|\hat{\theta}) + p \right] = E_{\theta^*} \left[ \Delta(\theta^*||\hat{\theta}) \right] + o_n(1)$$

Note that (under certain conditions)

$$E_{\theta^*} \left[ \Delta(\theta^*||\hat{\theta}) \right] = \Delta(\theta^*||\theta_0) + \frac{1}{2} (\hat{\theta} - \theta_0)' J(\theta_0) (\hat{\theta} - \theta_0) + o_n(1)$$
The aim is to show that

$$E_{\theta^*} \left[ L(y^n|\hat{\theta}) + p \right] = E_{\theta^*} \left[ \Delta(\theta^*||\hat{\theta}) \right] + o_n(1)$$

Note that (under certain conditions)

$$E_{\theta^*} \left[ \Delta(\theta^*||\hat{\theta}) \right] = \Delta(\theta^*||\theta_0) + \frac{1}{2} (\hat{\theta} - \theta_0)^' J(\theta_0)(\hat{\theta} - \theta_0) + o_n(1)$$

... and

$$\Delta(\theta^*||\theta_0) = E_{\theta^*} \left[ L(y^n|\hat{\theta}) \right] + \frac{1}{2} (\hat{\theta} - \theta_0)^' H(\hat{\theta})(\hat{\theta} - \theta_0) + o_n(1)$$
The aim is to show that

$$E_{\theta^*} \left[ L(y^n | \hat{\theta}) + p \right] = E_{\theta^*} \left[ \Delta(\theta^* || \hat{\theta}) \right] + o_n(1)$$

Note that (under certain conditions)

$$E_{\theta^*} \left[ \Delta(\theta^* || \hat{\theta}) \right] = \Delta(\theta^* || \theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)' J(\theta_0) (\hat{\theta} - \theta_0) + o_n(1)$$

... and

$$\Delta(\theta^* || \theta_0) = E_{\theta^*} \left[ L(y^n | \hat{\theta}) \right] + \frac{1}{2}(\hat{\theta} - \theta_0)' H(\hat{\theta}) (\hat{\theta} - \theta_0) + o_n(1)$$

Where

$$J(\theta_0) = \left[ \frac{\partial^2 \Delta(\theta^* || \theta)}{\partial \theta \partial \theta'} \right]_{\theta = \theta_0}, \quad H(\hat{\theta}) = \left[ \frac{\partial^2 L(y^n || \theta)}{\partial \theta \partial \theta'} \right]_{\theta = \hat{\theta}}$$
Since

\[ \frac{1}{2} E_{\theta^*} \left[ (\hat{\theta} - \theta_0)' J(\theta_0)(\hat{\theta} - \theta_0) \right] = \frac{p}{2} + o_n(1) \]

\[ \frac{1}{2} E_{\theta^*} \left[ (\hat{\theta} - \theta_0)' H(\hat{\theta})(\hat{\theta} - \theta_0) \right] = \frac{p}{2} + o_n(1) \]
Since
\[
\frac{1}{2}E_{\theta^*} \left[ (\hat{\theta} - \theta_0)' J(\theta_0)(\hat{\theta} - \theta_0) \right] = \frac{p}{2} + o_n(1)
\]
\[
\frac{1}{2}E_{\theta^*} \left[ (\hat{\theta} - \theta_0)' H(\hat{\theta})(\hat{\theta} - \theta_0) \right] = \frac{p}{2} + o_n(1)
\]

Then, substituting
\[
E_{\theta^*} \left[ \Delta(\theta^* || \hat{\theta}) \right] = \left( E_{\theta^*} \left[ L(y^n | \hat{\theta}) \right] + \frac{p}{2} + o_n(1) \right) + \frac{p}{2} + o_n(1)
\]
\[
= E_{\theta^*} \left[ L(y^n | \hat{\theta}) + p \right] + o_n(1)
\]
\[
\text{AIC}(\hat{\theta})
\]
References