Model Selection Tutorial #1: Akaike's Information Criterion

Daniel F. Schmidt and Enes Makalic

Melbourne, November 22, 2008

Content

Problem

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

$$\mathbf{y}^n \sim p^*$$

  where $\mathbf{y}^n \in \mathcal{Y}^n$.
- We wish to *learn* about p^* from $\mathbf{y}^n$.
- More precisely, we would like to discover the generating source p^*, or at least a *good* approximation of it, from nothing but $\mathbf{y}^n$

Statistical Models

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

$$p(\mathbf{y}^n|\boldsymbol{\theta}), \ \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

  where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is a *parameter* vector that indexes the particular model
- Such models satisfy

$$\int_{\mathbf{y}^n \in \mathcal{Y}^n} p(\mathbf{y}^n|\boldsymbol{\theta}) \mathrm{d}\mathbf{y}^n = 1$$

  for a *fixed* $\boldsymbol{\theta}$

Statistical Models ...

- An example would be the univariate normal distribution.

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

where

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

Content

Parameter Estimation

- Given a statistical model and data $\mathbf{y}^n$, we would like to take a guess at a plausible value of θ
- 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

Method of Maximum Likelihood (ML), Part 1

- A heuristic procedure introduced by R. A. Fisher
- Possesses good properties in many cases
- Is very general and easy to understand
- To estimate parameters θ for a statistical model from $\mathbf{y}^n$, solve

$$\hat{\theta}(\mathbf{y}^n) = \arg\max_{\theta \in \Theta} \{p(\mathbf{y}^n | \theta)\}$$

or, more conveniently

$$\hat{\theta}(\mathbf{y}^n) = \arg\min_{\theta \in \Theta} \{-\log p(\mathbf{y}^n | \theta)\}$$

Method of Maximum Likelihood (ML), Part 2

- Example : Estimating the mean parameter μ of a univariate normal distribution
- Negative log-likelihood function :

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

Method of Maximum Likelihood (ML), Part 2

- Example : Estimating the mean parameter μ 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 μ yields

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

Method of Maximum Likelihood (ML), Part 2

- Example : Estimating the mean parameter μ 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 μ 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 μ yields

$$\hat{\mu}(\mathbf{y}^n) = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Univariate Polynomial Regression

- 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 τ 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 $\boldsymbol{\theta}^{(k)} = (\tau, \beta_0, \ldots, \beta_k)$.

Univariate Polynomial Regression

- A more complex model : k-order polynomial regression
- Let each $y(x)$ be distributed as per a univariate normal with variance τ 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 $\boldsymbol{\theta}^{(k)} = (\tau, \beta_0, \ldots, \beta_k)$.
- In this model the data $\mathbf{y}^n$ is associated with a $\mathbf{x}^n$ which are *known*

Univariate Polynomial Regression

- A more complex model : k-order polynomial regression
- Let each $y(x)$ be distributed as per a univariate normal with variance τ 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 $\boldsymbol{\theta}^{(k)} = (\tau, \beta_0, \ldots, \beta_k)$.

- In this model the data $\mathbf{y}^n$ is associated with a $\mathbf{x}^n$ which are *known*
- *Given* an order k, maximum likelihood can be used to estimate $\boldsymbol{\theta}^{(k)}$

Univariate Polynomial Regression

- A more complex model : k-order polynomial regression
- Let each $y(x)$ be distributed as per a univariate normal with variance τ 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 $\boldsymbol{\theta}^{(k)} = (\tau, \beta_0, \ldots, \beta_k)$.
- In this model the data $\mathbf{y}^n$ is associated with a $\mathbf{x}^n$ which are *known*
- *Given* an order k, maximum likelihood can be used to estimate $\boldsymbol{\theta}^{(k)}$
- 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 $\hat{\beta}^{(k)}(\mathbf{y}^n)$ to minimise

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

This is called the *residual variance*.

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 $\hat{\beta}^{(k)}(\mathbf{y}^n)$ to minimise

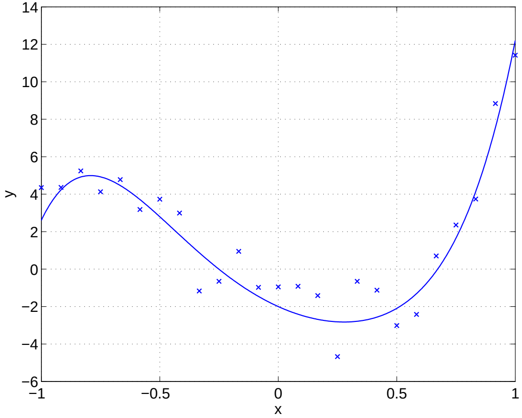
$$\hat{\tau}^{(k)}(\mathbf{y}^n) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\mu}^{(k)}(x_i) \right)^2$$

  This is called the *residual variance*.
- The likelihood function $L(\mathbf{y}^n | \hat{\theta}^{(k)}(\mathbf{y}^n))$ made by plugging in the Maximum Likelihood estimates is

$$L(\mathbf{y}^n | \hat{\theta}^{(k)}(\mathbf{y}^n)) = \frac{n}{2} \log \left(2\pi \hat{\tau}^{(k)}(\mathbf{y}^n) \right) + \frac{n}{2}$$

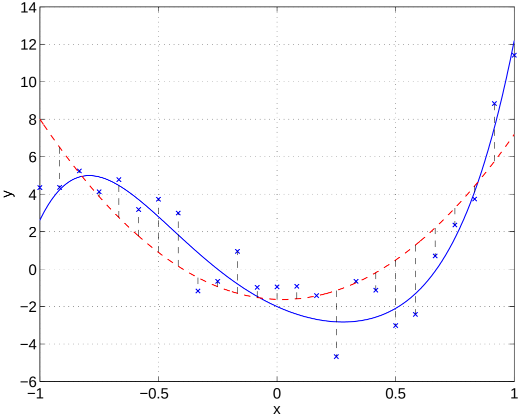
Method of Maximum Likelihood (ML), Part 4

'Truth' : $\mu(x) = 9.7x^5 + 0.8x^3 + 9.4x^2 - 5.7x - 2, \tau = 1$



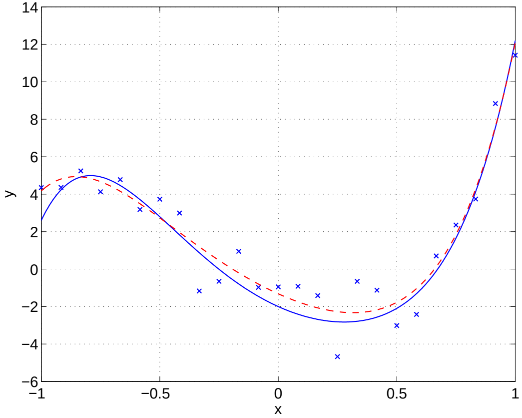
Method of Maximum Likelihood (ML), Part 4

Polynomial fit, $k = 2$, $\hat{\tau}^{(2)}(\mathbf{y}) = 4.6919$



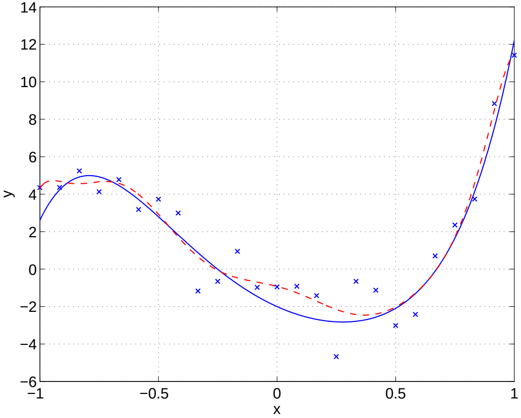
Method of Maximum Likelihood (ML), Part 4

Polynomial fit, $k = 5$, $\hat{\tau}^{(5)}(\mathbf{y}) = 1.1388$



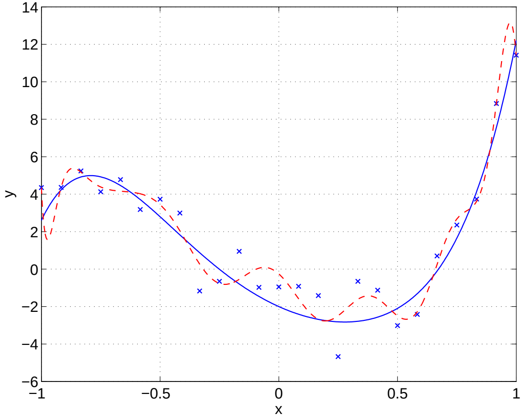
Method of Maximum Likelihood (ML), Part 4

Polynomial fit, $k = 10$, $\hat{\tau}^{(10)}(\mathbf{y}) = 1.0038$



Method of Maximum Likelihood (ML), Part 4

Polynomial fit, $k = 20$, $\hat{\tau}^{(20)}(\mathbf{y}) = 0.1612$



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

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

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

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

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

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

Content

Kullback-Leibler Divergence

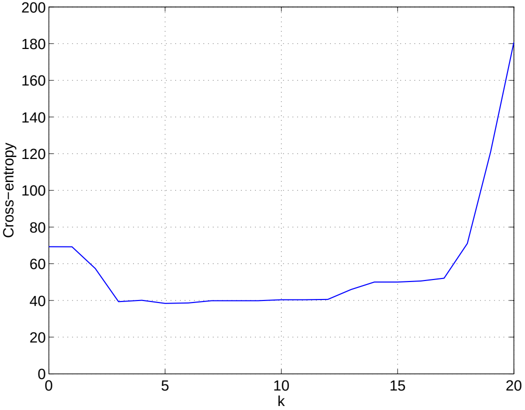
- AIC is based on estimating the Kullback-Leibler (KL) divergence
- The Kullback-Leibler divergence

$$KL(f||g) = \underbrace{-\int_{\mathcal{Y}^n} f(\mathbf{y}^n) \log g(\mathbf{y}^n) d\mathbf{y}^n}_{Cross-entropy} + \underbrace{\int_{\mathcal{Y}^n} f(\mathbf{y}^n) \log f(\mathbf{y}^n) d\mathbf{y}^n}_{Entropy}$$

- Cross-entropy, $\Delta(f||g)$, is the ‘expected negative log-likelihood’ of data coming from f under g

Kullback-Leibler Divergence

- 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 !

Akaike's Information Criterion

- The AIC score for a model is

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

  where p is the number of free model parameters.
- Using AIC one chooses the model that solves

$$\hat{k} = \underset{k\in\{0,1,\ldots\}}{\arg\min} \left\{ \mathrm{AIC}(\hat{\theta}^{(k)}(\mathbf{y}^n)) \right\}$$

Properties of AIC

- Under certain conditions the AIC score satisfies

$$\mathrm{E}_{\boldsymbol{\theta}^*}\left[\mathrm{AIC}(\hat{\boldsymbol{\theta}}(\mathbf{y}^n))\right] = \mathrm{E}_{\boldsymbol{\theta}^*}\left[\Delta(\boldsymbol{\theta}^*||\hat{\boldsymbol{\theta}}(\mathbf{y}^n))\right] + o_n(1)$$

  where $o_n(1) \to 0$ as $n \to \infty$
- In words, the AIC score is an *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*

Conditions for AIC to apply

- 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 $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$

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 $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$
- Every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ must map to a unique distribution $p(\cdot|\boldsymbol{\theta})$

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 $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$
- Every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ must map to a unique distribution $p(\cdot|\boldsymbol{\theta})$
- The Maximum Likelihood estimates must be consistent and be approximately normally distributed for large n

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 $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$
- Every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ must map to a unique distribution $p(\cdot|\boldsymbol{\theta})$
- The Maximum Likelihood estimates must be consistent and be approximately normally distributed for large n
- $L(\boldsymbol{\theta})$ must be twice differentiable with respect to $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

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 θ map to the same distribution

When not to use AIC

- Multilayer Perceptron Neural Networks
  - Many different θ 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 θ 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

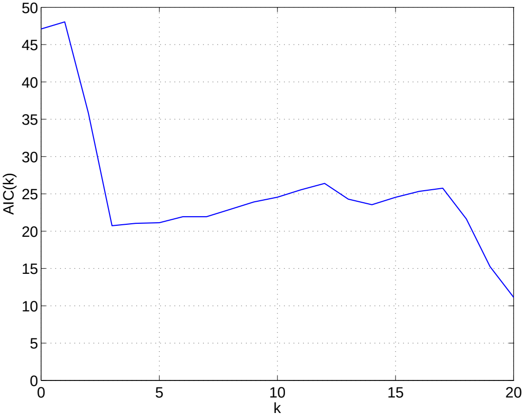
When not to use AIC

- Multilayer Perceptron Neural Networks
  - Many different θ 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

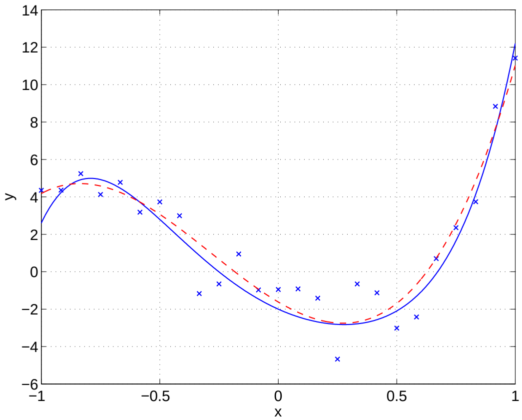
- AIC criterion for polynomials

$$\mathrm{AIC}(k) = \frac{n}{2} \log 2\pi \hat{\tau}^{(k)}(\mathbf{y}^n) + \frac{n}{2} + (k+2)$$



Application to polynomials

- AIC selects $\hat{k} = 3$



Improvements to AIC

- For some model types it is possible to derive improved estimates of the cross-entropy
- Under certain conditions, the 'corrected' AIC (AICc) criterion

$$\mathrm{AIC}_c(\hat{\theta}(\mathbf{y}^n)) = -\log p(\mathbf{y}^n|\hat{\theta}(\mathbf{y}^n)) + \frac{n(p+1)}{n-p-2}$$

satisfies

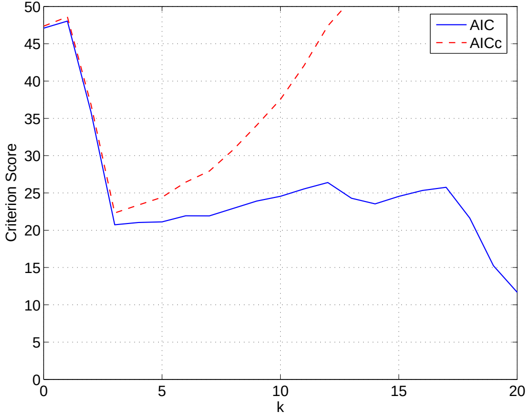
$$\mathrm{E}_{\theta^*}\left[\mathrm{AIC}_c(\hat{\theta}(\mathbf{y}^n))\right] = \mathrm{E}_{\theta^*}\left[\Delta(\theta^*||\hat{\theta}(\mathbf{y}^n))\right]$$

- In words, it is an exactly unbiased estimator of the cross-entropy, even for finite n

Application to polynomials

- AICc criterion for polynomials

$$\mathrm{AIC}_c(k) = \frac{n}{2} \log 2\pi\hat{\tau}^{(k)}(\mathbf{y}^n) + \frac{n}{2} + \frac{n(k+2)}{n-k-3}$$



Using AICc

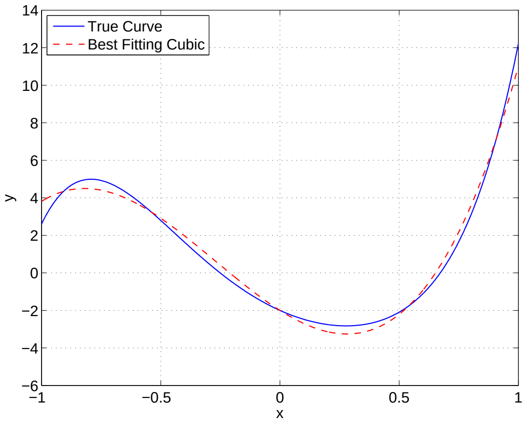
- Tends to perform better than AIC, especially when n/p is *small*
- Theoretically only valid for homoskedastic *linear models*; these include
  - Linear regression models, including linear function approximation
  - Autoregressive Moving Average (ARMA) models
  - Linear smoothers (kernel, local regression,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

Content

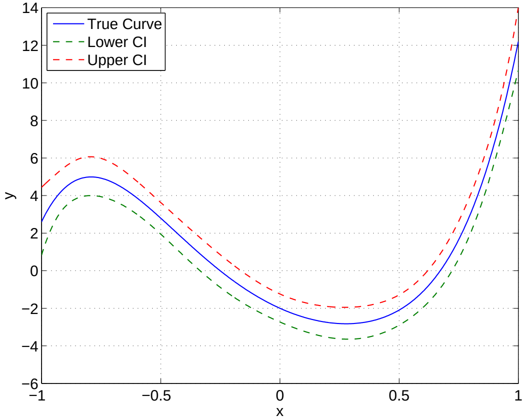
Some theory

- 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



Discrepancy due to Estimation



Derivation

- The aim is to show that

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

Derivation

- The aim is to show that

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

- Note that (under certain conditions)

$$\mathrm{E}_{\boldsymbol{\theta}^*}\left[\Delta(\boldsymbol{\theta}^*||\hat{\boldsymbol{\theta}})\right] = \Delta(\boldsymbol{\theta}^*||\boldsymbol{\theta}_0) + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{J}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_n(1)$$

Derivation

- The aim is to show that

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

- Note that (under certain conditions)

$$\mathrm{E}_{\boldsymbol{\theta}^*}\left[\Delta(\boldsymbol{\theta}^*||\hat{\boldsymbol{\theta}})\right] = \Delta(\boldsymbol{\theta}^*||\boldsymbol{\theta}_0) + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{J}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_n(1)$$

- ... and

$$\Delta(\boldsymbol{\theta}^*||\boldsymbol{\theta}_0) = \mathrm{E}_{\boldsymbol{\theta}^*}\left[L(\mathbf{y}^n|\hat{\boldsymbol{\theta}})\right] + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{H}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_n(1)$$

Derivation

- The aim is to show that

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

- Note that (under certain conditions)

$$\mathrm{E}_{\boldsymbol{\theta}^*}\left[\Delta(\boldsymbol{\theta}^*||\hat{\boldsymbol{\theta}})\right] = \Delta(\boldsymbol{\theta}^*||\boldsymbol{\theta}_0) + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{J}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_n(1)$$

- ... and

$$\Delta(\boldsymbol{\theta}^*||\boldsymbol{\theta}_0) = \mathrm{E}_{\boldsymbol{\theta}^*}\left[L(\mathbf{y}^n|\hat{\boldsymbol{\theta}})\right] + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{H}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_n(1)$$

- Where

$$\mathbf{J}(\boldsymbol{\theta}_0) = \left[\left.\frac{\partial^2 \Delta(\boldsymbol{\theta}^*||\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right], \quad \mathbf{H}(\hat{\boldsymbol{\theta}}) = \left[\left.\frac{\partial^2 L(\mathbf{y}^n||\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]$$

Derivation

- Since

$$
\begin{aligned}
\frac{1}{2}\mathrm{E}_{\boldsymbol{\theta}^*}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{J}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right] &= \frac{p}{2} + o_n(1) \\
\frac{1}{2}\mathrm{E}_{\boldsymbol{\theta}^*}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{H}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right] &= \frac{p}{2} + o_n(1)
\end{aligned}
$$

Derivation

- Since

$$\begin{aligned}\frac{1}{2}\mathrm{E}_{\boldsymbol{\theta}^*}\left[(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)'\mathbf{J}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right] &= \frac{p}{2}+o_n(1) \\ \frac{1}{2}\mathrm{E}_{\boldsymbol{\theta}^*}\left[(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)'\mathbf{H}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right] &= \frac{p}{2}+o_n(1)\end{aligned}$$

- Then, substituting

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\theta}^*}\left[\Delta(\boldsymbol{\theta}^*||\hat{\boldsymbol{\theta}})\right] &= \left(\mathrm{E}_{\boldsymbol{\theta}^*}\left[L(\mathbf{y}^n|\hat{\boldsymbol{\theta}})\right]+\frac{p}{2}+o_n(1)\right)+\frac{p}{2}+o_n(1) \\ &= \mathrm{E}_{\boldsymbol{\theta}^*}\underbrace{\left[L(\mathbf{y}^n|\hat{\boldsymbol{\theta}})+p\right]}_{\mathrm{AIC}(\hat{\boldsymbol{\theta}})}+o_n(1)\end{aligned}$$

Content

References

- S. Kullback and R. A. Leibler, ‘On Information and Sufficiency’, *The Annals of Mathematical Statistics*, Vol. 22, No. 1, pp. 79–86, 1951
- H. Akaike, ‘A new look at the statistical model identification’, *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, pp. 716–723, 1974
- H. Linhart and W. Zucchini, *Model Selection*, John Wiley and Sons, 1986
- C. M. Hurvich and C. Tsai, ‘Regression and Time Series Model Selection in Small Samples’, *Biometrika*, Vol. 76, pp. 297–307, 1989
- J. E. Cavanaugh, ‘Unifying the Deriviations for the Akaike and Corrected Akaike Information Criteria’, *Statistics & Probability Letters*, Vol. 33, pp. 201–208, 1997