The Behaviour of the Akaike Information Criterion when Applied to Non-nested Sequences of Models

Daniel F. Schmidt and Enes Makalic

Centre for Molecular, Environmental, Genetic & Analytic (MEGA) Epidemiology
School of Population Health
University of Melbourne

23rd Australasian Joint Conference on Artificial Intelligence
8th of December 2010

Content

Overview of Paper

- Model selection, i.e., choosing between models of different complexity to fit the data $\mathbf{y}$
- Examine the use of the Akaike Information Criterion (AIC) in the case of non-nested model selection
- We derive a novel upper-bound on the bias of the AIC
- Crucially, this bound does not depend on sampe sizle n
  - Depends on the number of models under consideration

$\Rightarrow$ The bias cannot be overcome even as $n \to \infty$

Model Selection

- We have observed data $\mathbf{y} = (y_1, \ldots, y_n)$
- Have a candidate set, Γ, of models to explain the data
  - Some models more complex than others
  - Want to find model that justifiably best fits the data
- More complex models (more parameters) always fit better
  - Is the extra fit warranted, or due to random variation ?
- Use model selection criterion to choose between models
  - One method is Akaike Information Criterion (AIC)

Example Model : Polynomial Regression

- Consider data in pairs (x_i, y_i)
  - Want to made predictions about y_i given x_i
- One possible model is polynomial regression

$$\mathrm{E}\left[y_i | x_i\right] = \sum_{j=0}^{q} \beta_j x_i^j$$

- Number of terms/maximum order controls complexity
  - e.g., a quadratic model (3 parameters)

$$\mathrm{E}\left[y_i | x_i\right] = \beta_2 x_i^2 + \beta_1 x_i + \beta_0$$

always fits better than a linear model (2 parameters)

$$\mathrm{E}\left[y_i | x_i\right] = \beta_1 x_i + \beta_0$$

- But is the extra fit just “learning” noise ?

Nested Models

- We can partition the complete set of models Γ into subsets

$$\Gamma = \Gamma_1 \cup \Gamma_2 \ldots ...$$

  where Γ_j is the set of all models with j free parameters
- Nested sequences of models play an important role
- A sequence of models is nested iff :
  - There is only one model with j parameters, for all j
  - Models with k parameters can exactly represent all models with $j < k$ parameters

Nested Models, Example

- To return to polynomial regression ...
  - Frame the problem of model selection as order selection.
  - Assume maximum order is cubic

$$\Gamma_0 = \{\}, \Gamma_1 = \{1\}, \Gamma_2 = \{x, 1\}, \Gamma_3 = \{x^2, x, 1\}, \Gamma_4 = \{x^3, x^2, x, 1\}$$

  $\Rightarrow$ The above forms a nested sequence :
- By setting relevant β coefficients to zero ...
  - A linear model can represent a constant model
  - A quadratic can represent a linear and a constant model
  - A cubic model can represent a quadratic, linear and constant model
  - All models can represent the empty model (no coefficients)
- e.g., a cubic with $\beta_3 = \beta_2 = 0$ is effectively a linear model

Non-Nested Models, Example

- Alternatively frame the problem as selecting individual polynomial terms

$$
\begin{aligned}
\Gamma_0 &= \{\} \\
\Gamma_1 &= \{1\}, \{x\}, \{x^2\}, \{x^3\} \\
\Gamma_2 &= \{x, 1\}, \{x^2, 1\}, \{x^3, 1\}, \{x^2, x\}, \{x^3, x\}, \{x^3, x^2\} \\
\Gamma_3 &= \{x^2, x, 1\}, \{x^3, x, 1\}, \{x^3, x^2, 1\}, \{x^3, x^2, x\} \\
\Gamma_4 &= \{x^3, x^2, x, 1\}
\end{aligned}
$$

  $\Rightarrow$ The above is *not* a nested model class. There are ...

  - 4 models with 1 parameter
  - 6 models with 2 parameters
  - 4 models with 3 parameters

- e.g., model $\{x^3, x^2\}$ *cannot* represent models $\{1\}$ or $\{x\}$

Statistical Models

- The word "model" denotes a conditional probability distribution, $p(\mathbf{y}|\boldsymbol{\theta}_\gamma)$
  - γ is a symbol denoting a model from the set Γ
  - $\boldsymbol{\theta}_\gamma \in \Theta_\gamma$ are the parameters associated with the model
- e.g., for polynomial regression, assuming normal noise and $\gamma = \{x^2, 1\}$ specifies

$$y_i|x_i, \beta_2, \beta_0, \sigma^2 \sim N\left(\beta_2 x_i^2 + \beta_0, \sigma^2\right)$$

where $\boldsymbol{\theta}_\gamma = (\beta_2^2, \beta_0, \sigma^2)$, and $N(\cdot)$ is a normal distribution

The “true” model

- The final concept we need is that of the “true” model
- If we let
  - $\gamma_* \in \Gamma$ denote the true model
  - $\boldsymbol{\theta}_*$ denote the true parameters in model γ_*
- By assuming a “true” model, we are assuming the data is generated by the conditional distribution represented by this model
- This is an important assumption for many model selection criterion

Content

Fitting Models

- Use maximum likelihood to fit models to data

$$\hat{\boldsymbol{\theta}}_\gamma = \underset{\boldsymbol{\theta}_\gamma \in \Theta_\gamma}{\arg\max} \{p(\mathbf{y}|\boldsymbol{\theta}_\gamma)\}$$

- Need to measure how close a fitted model is to the truth
  - One possibility is Kullback–Leibler (KL) divergence
  - Let

$$d(\boldsymbol{\theta}_*, \hat{\boldsymbol{\theta}}_\gamma) = 2\mathrm{E}_{\boldsymbol{\theta}_*}\left[\log 1/p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma)\right]$$

  - The KL divergence is given by

$$2\Delta(\boldsymbol{\theta}_*||\hat{\boldsymbol{\theta}}_\gamma) = \underbrace{d(\boldsymbol{\theta}_*, \hat{\boldsymbol{\theta}}_\gamma)}_{cross\ entropy} - \underbrace{d(\boldsymbol{\theta}_*, \boldsymbol{\theta}_*)}_{entropy}$$

- Ideally choose the model with the smallest KL divergence
  $\Rightarrow$ Relies on knowledge of the truth

Akaike's Information Criterion (AIC)

- Akaike's idea was to *estimate* the expected KL divergence for each model
  $\Rightarrow$ Select model with lowest expected KL divergence
- Select model that minimises

$$\mathrm{AIC}(\gamma) = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma) + 2k_\gamma$$

  where k_γ is the number of free parameters possessd by model γ
- Under certain conditions (model γ contains truth, nested models)

$$\mathrm{E}\left[\mathrm{AIC}(\gamma)\right] = \mathrm{E}\left[d(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_\gamma)\right] + o_n(1)$$

Content

Main Result 1

- Main result concerns non-nested model sequences
- Assume usual AIC regularity conditions
- Setup
  - Starting from the true model γ_* with no parameters
  - Choosing between candidate models with k parameters
  - Let

$$a_m = \log p(\mathbf{y}|\boldsymbol{\theta}_*) - \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_m), \quad m \in \Gamma_k$$

    be the improvement in fit for the candidate models.
  - We assume the a_m are all independent random variates
- Selecting best k model parameter is equivalent to

$$\hat{m}_k = \underset{m \in \Gamma_k}{\arg\max}\{a_m\}$$

Main Result 2

Theorem

$$2\mathrm{E}_{\boldsymbol{\theta}^*}\left[\log 1/p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\hat{m}_k})\right] = 2\alpha(k, q_k) = \mathrm{E}_{\boldsymbol{\theta}^*}\left[d(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_{\hat{m}_k})\right] + o_n(1)$$

where

$$\alpha(k, q_k) = \mathrm{E}_{\chi^2_k}\left[\max\left\{z_1, \ldots, z_{q_k}\right\}\right]$$

and

- $z_1, \ldots, z_{q_k}$ *are iid* χ^2_k *variates with* k *degrees of freedom*
- $q_k = |\Gamma_k|$ *is the number of candidate models with* k *parameters*

Implications 1

- What does it mean ?
- If model class is *nested*, $q_k = 1$ and $\alpha(k, q_k) = k$
  $\Rightarrow$ we recover the regular AIC
- If $q_k > 1$, $\alpha(k, q_k) > k$
  - regular AIC underestimates expected cross-entropy
  - as $\alpha(k, q_k)$ does not depend on n, this bias will remain even as $n \to \infty$

  $\Rightarrow$ Increased likelihood of overfitting depends on q_k !

Implications 2

- How serious is it ?
- Consider a simple regression setup
  - We have q orthogonal covariates

$$\Gamma_1 = \{1\}, \{2\}, \ldots, \{q\}$$

  - The "true" model includes no covariates, i.e., $\gamma_* = \{\}$
  - Use AIC to decide whether to prefer a model in Γ_1 to the true, empty model γ_*
- Using our Theorem, we can determine probability of overfitting when using AIC

Implications 2, cont'd

TABLE: Probability of Overfitting

q_1	P(overfit)
1	0.157
2	0.290
3	0.402
4	0.496
5	0.575
8	0.746
10	0.819
15	0.923
25	0.986
50	0.999
100	1.000

Implications 3

- What if the a_m variates are not independent ?
  - e.g., if several regression models include the same covariates their a_m variates will be correlated
- In this case our Theorem acts as an upper-bound
- How tight this bound is in many usual model selection settings is a topic for future research

Implications 4

- Is the result entirely negative ?
- It is possible to use the main result to construct improved "AIC-like" criteria
  - Our paper describes one such procedure for forward selection of features in regression models
- The basic procedure :
  - if all q remaining features are noise, expected improvement in fit due to including a noise feature is

$$\mathrm{E}\left[a_m\right] = (1/2)\alpha(q, 1)$$

  - Can be used to choose a threshold for inclusion of a feature
- Applied with success to denoising in our paper

Conclusion

- Examined the use of the Akaike Information Criterion (AIC) in the case of non-nested model selection
- We derive a novel upper-bound on the bias of the AIC
- Crucially, this bound does not depend on sampe sizle n
  - Depends on the number of models under consideration

$\Rightarrow$ The bias cannot be overcome even as $n \to \infty$

References

- Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, Vol. 19, pp. 716–723
- Kullback, S., Leibler, R.A. On information and sufficiency. *The Annals of Mathematical Statistics*, 1951, Vol. 22, pp. 79–86
- Linhart, H., Zucchini, W. Model Selection. Wiley, New York, 1986