Derivations of AICc and GCVc
CVEK-boot

Wenying Deng

April 21, 2018

Date Performed:	April 8, 2018
Instructor:	Jeremiah.Liu

1 Derivation of AICc

1.1 From KL info to AIC[1]

Consider the situation where $x_1, x_2, ..., x_n$ are obtained as the results of n independent observations of a random variable with pdf $g(x)$. If a parametric family of density function is given by $f(y \mid \theta)$ with a vector parameter θ, the average log-likelihood is given by

$$\frac{1}{n}\sum_{i=1}^{n} log f(x_i \mid \theta)$$

As n increases, this average tends, with probability 1, to

$$S(g; f(\cdot \mid \theta)) = \int g(x) log f(x \mid \theta) dx$$

The difference

$$I(g; f(\cdot \mid \theta)) = S(g; g) - S(g; f(\cdot \mid \theta)) \tag{1}$$

is known as the Kullback-Leibler mean information for discrimination between $g(x)$ and $f(x \mid \theta)$ and takes positive value, unless $f(x \mid \theta) = g(x)$ holds almost everywhere.

Consider the situation where $g(x) = f(x \mid \theta_0)$. For this case $I(g; f(\cdot \mid \theta))$and $S(g; f(\cdot \mid \theta))$ will simply be denoted by $I(\theta_0; \theta)$ and $S(\theta_0; \theta)$, respectively. When θ is sufficiently close to θ_0, $I(\theta_0; \theta)$ admits an approximation

$$I(\theta_0; \theta_0 + \Delta\theta) = \frac{1}{2} \parallel \Delta\theta \parallel_J^2$$

where $\| \Delta\theta \|_J^2 = \Delta\theta^T J \Delta\theta$ and J is the Fisher information matrix which is positive definite and defined by

$$J_{ij} = E\{\frac{\partial logf(X \mid \theta)}{\partial\theta_i}\frac{\partial logf(X \mid \theta)}{\partial\theta_j}\}$$

When the MLE $\hat{\theta}$ of θ_0 lies very close to θ_0, the deviation of the distribution defined by $f(x \mid \theta)$ from the true distribution $f(x \mid \theta_0)$ in terms of the variation of $S(g; f(\cdot \mid \theta))$ will be measured by $\frac{1}{2} \| \theta - \theta_0 \|_J^2$. Consider the situation where the variation of θ for maximizing the likelihood is restricted to a lower dimensional subspace Θ of θ which does not include θ_0. For the MLE $\hat{\theta}$ of θ_0 restricted in Θ, if θ which is in Θ and gives the maximum of $S(\theta_0; \theta)$ is sufficiently close to θ_0, it can be shown that the distribution of $n \| \hat{\theta} - \theta \|_J^2$ for sufficiently large n is approximately under certain regularity conditions by a chi-square distribution with the df equal to the dimension of the restricted parameter space. Thus,

$$\begin{aligned} n \| \hat{\theta} - \theta_0 \|_J^2 &= n \| \hat{\theta} - \theta + \theta - \theta_0 \|_J^2 \\ &= n \| \hat{\theta} - \theta \|_J^2 + n \| \theta_0 - \theta \|_J^2 \end{aligned}$$

$$E[2nI(\theta_0; \hat{\theta})] = n \| \theta_0 - \theta \|_J^2 + k \tag{2}$$

where k is the dimension of Θ or the number of parameters independently adjusted for the maximization of the likelihood.

Why equation(2) becomes $AIC = 2k - 2ln(\hat{L})$ like the formula we see in wiki?
The relation(2) is based on the fact that the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$ is approximated by a Gaussian distribution with mean zero and variance matrix J^{-1}.
Let's expand the log-likelihood $\frac{1}{n}\sum_{i=1}^n logf(x_i \mid \hat{\theta})$ at θ_0:

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n logf(x_i \mid \hat{\theta}) \\ \approx&\frac{1}{n}\sum_{i=1}^n logf(x_i \mid \theta_0) + (\hat{\theta} - \theta_0)^T[\frac{1}{n}\sum_{i=1}^n \frac{\partial logf(x_i \mid \theta)}{\partial\theta}] \mid_{\theta=\theta_0} \\ &+ \frac{1}{2}(\hat{\theta} - \theta_0)^T[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 logf(x_i \mid \theta)}{\partial\theta^2}] \mid_{\theta=\theta_0} (\hat{\theta} - \theta_0) \\ \approx&\frac{1}{n}\sum_{i=1}^n logf(x_i \mid \theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^T[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2} logf(x_i \mid \theta)] \mid_{\theta=\theta_0} (\hat{\theta} - \theta_0) \end{aligned}$$

Since,

$$
\begin{aligned}
&[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial logf(x_i \mid \theta)}{\partial\theta}] \mid_{\theta=\theta_0} \approx E[\frac{\partial logf(x \mid \theta)}{\partial\theta} \mid_{\theta=\theta_0}] = 0 \\
&[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}logf(x_i \mid \theta)] \mid_{\theta=\theta_0} \approx - I(\theta_0)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&2[\sum_{i=1}^{n} logf(x_i \mid \theta_0) - \sum_{i=1}^{n} logf(x_i \mid \hat{\theta})] + k \\
&\approx (\theta_0 - \hat{\theta})^T \sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}logf(x_i \mid \theta)(\theta_0 - \hat{\theta}) + k \\
&\approx n \parallel \theta_0 - \theta \parallel_J^2
\end{aligned}
$$

The reason why I add k to $2[\sum_{i=1}^{n} logf(x_i \mid \theta_0) - \sum_{i=1}^{n} logf(x_i \mid \hat{\theta})]$ is we need a correction for the bias introduced by replacing θ by $\hat{\theta}$. Therefore, (2) becomes

$$
\begin{aligned}
E[2nI(\theta_0;\hat{\theta})] =& 2k + 2[\sum_{i=1}^{n} logf(x_i \mid \theta_0) - \sum_{i=1}^{n} logf(x_i \mid \hat{\theta})] \\
=& 2k - 2log(\hat{L}) + 2\sum_{i=1}^{n} logf(x_i \mid \theta_0)
\end{aligned}
$$

Moreover, since we are optimizing with respect to $\hat{\theta}$, we don't need to consider $2\sum_{i=1}^{n} logf(x_i \mid \theta_0)$, thus giving our objective fuction:

$$AIC = 2k - 2ln(\hat{L}) \tag{3}$$

1.2 From AIC to AICc[2]

Now we focus on minimizing $\Delta(\theta, \sigma^2) = -S(g; f(\cdot \mid \theta, \sigma^2))$, taking into account σ^2 as a parameter. Suppose the true g corresponds to true model μ

$$y = \mu + \epsilon \tag{4}$$

where $\epsilon \sim N(0, \sigma_0^2)$.
And the estimating f corresponds to the approximating model $h(\theta)$

$$y = h(\theta) + u \tag{5}$$

where $u \sim N(0, \sigma^2)$.
We have

$$
\begin{aligned}
\Delta(\theta, \sigma^2) =& - 2E_g log\{(2\pi\sigma^2)^{-\frac{1}{2}n} exp[-\{y - h(\theta)\}^T\{y - h(\theta)\}/(2\sigma^2)]\} \\
=& nlog(2\pi\sigma^2) + E_g\{\mu + \epsilon - h(\theta)\}^T\{\mu + \epsilon - h(\theta)\}/\sigma^2 \\
=& nlog(2\pi\sigma^2) + n\sigma_0^2/\sigma^2 + \{\mu - h(\theta)\}^T\{\mu - h(\theta)\}/\sigma^2
\end{aligned}
$$

A reasonable criterion for judging the quality of the approximating family in the light of the data is $E_g\{\Delta(\hat{\theta}, \hat{\sigma}^2)\}$, where $\hat{\theta}$ and $\hat{\sigma}^2$ are the MLE: $\hat{\theta}$ minimizes $\{y - h(\theta)\}^T\{y - h(\theta)\}$ and

$$\hat{\sigma}^2 = \{y - h(\hat{\theta})\}^T\{y - h(\hat{\theta})\}/n$$

Ignoring the constant $nlog(2\pi)$, we have

$$\Delta(\hat{\theta}, \hat{\sigma}^2) = nlog(\hat{\sigma}^2) + n\sigma_0^2/\hat{\sigma}^2 + \{\mu - h(\hat{\theta})\}^T\{\mu - h(\hat{\theta})\}/\hat{\sigma}^2$$

Consider the equation(3) we derived just now, in this case,

$$\begin{aligned}
AIC =& 2(k+1) - 2ln(\hat{L}) \\
=& 2(k+1) - 2[-\frac{n}{2}log(\hat{\sigma}^2) - \frac{n}{2}log(2\pi) - \frac{1}{2\hat{\sigma}^2}\{y - h(\hat{\theta})\}^T\{y - h(\hat{\theta})\}] \\
=& 2(k+1) + nlog(\frac{1}{n}\{y - h(\hat{\theta})\}^T\{y - h(\hat{\theta})\}) + n + nlog(2\pi)
\end{aligned}$$

where k becomes $k+1$ due to the fact that we explicitly estimate σ^2 here. Again, ignoring the constant $nlog(2\pi)$ and plugging $\hat{\sigma}^2 = \{y - h(\hat{\theta})\}^T\{y - h(\hat{\theta})\}/n$ in, we have

$$AIC = n(log\hat{\sigma}^2 + 1) + 2(k+1) \tag{6}$$

Now assume that the approximating models include the true one. In this case, the mean response function μ of the true model can be written as $\mu = h(\theta_0)$, where θ_0 is an $k \times 1$ vector. The linear expansion of $h(\hat{\theta})$ at $\theta = \theta_0$ is give by

$$h(\hat{\theta}) \approx h(\theta_0) + V(\hat{\theta} - \theta_0)$$

where $V = \frac{\partial h}{\partial \theta}$ evaluated at $\theta = \theta_0$. Then under the true model

$$\hat{\theta} - \theta_0 \approx N(0, \sigma_0^2(V^TV)^{-1}) \tag{7}$$

the quantity $n\hat{\sigma}^2/\sigma_0^2$ is approximately distributed as χ^2_{n-k} independently of $\hat{\theta}$, and

$$\begin{aligned}
&(\frac{n-m}{nm})\frac{1}{\hat{\sigma}^2}\{\mu - h(\hat{\theta})\}^T\{\mu - h(\hat{\theta})\} \\
=&(\frac{n-m}{nm})\frac{1}{\hat{\sigma}^2}\{h(\theta_0) - h(\hat{\theta})\}^T\{h(\theta_0) - h(\hat{\theta})\} \\
\approx&(\frac{n-m}{nm})\frac{1}{\hat{\sigma}^2}(\hat{\theta} - \theta_0)^TV^TV(\hat{\theta} - \theta_0)
\end{aligned}$$

is approximately distributed as $F(m, n-m)$. Thus,

$$E_g\{\Delta(\hat{\theta}, \hat{\sigma}^2)\} \approx E_g(nlog(\hat{\sigma}^2)) + \frac{n^2}{n-m-2} + \frac{nm}{n-m-2}$$

Consequently, we obtain

$$AICc = nlog(\hat{\sigma}^2) + n\frac{1 + m/n}{1 - (m+2)/n} \tag{8}$$

1.3 From AICc to paper.2015[3]

Back to the 21st century paper.
Suppose we have data $\{\mathbf{y}, \mathbf{x}\}$, comprising n observations of a continuous outcome Y and p covariates $\mathbf{X}$, with the covariate matrix $\mathbf{x}$ regarded as fixed. We relate Y and $\mathbf{X}$ by a linear model, $E(Y) = \beta_0 + \mathbf{X^T}\beta$, with the errors distribute as normal distribution.
Note: p here doesn't contain the coefficient β_0 and σ^2, therefore, corresponding to Part **1.2**, $p + 2 = k + 1$.

$$(equation(6) - n)/n, k + 1 \to p + 2 \Rightarrow equation(2.5) inthepaper$$
$$(equation(8) - n)/n, k + 1 \to p + 2 \Rightarrow equation(2.7) inthepaper$$

2 Derivation of GCVc[3]

According to equation(2.4) in the paper,

$$\lambda_{GCV} = argmin_{\lambda}\{logy^T(I_n - P_\lambda)^2 y - 2log(1 - \frac{Trace(P_\lambda)}{n} - \frac{1}{n})\} \quad (9)$$

The difference between the objective function we derived in winter break and this one is the extra term $-\frac{1}{n}$. This is because at that time we assume β_0 is known, while now we need to re-estimate β_0 every time we re-centering y at each fold.

The motivation of GCVc is to take σ^2 into account, thus subtracting one more $1/n$ term:

$$\lambda_{GCVc} = argmin_{\lambda}\{logy^T(I_n - P_\lambda)^2 y - 2log(1 - \frac{Trace(P_\lambda)}{n} - \frac{2}{n})_+\} \quad (10)$$

When fitting GCVc, the effective number of remaining parameters is less than $n - 2$, and perfect fit of the observations to the predictions, given by $\lambda = 0$, cannot occur.

3 References

1. http://ieeexplore.ieee.org/document/1100705/

2. https://academic.oup.com/biomet/article-abstract/76/2/297/265326?redirectedFrom=fulltext

3. https://www.ncbi.nlm.nih.gov/pubmed/26985140