Model Selection — Akaike Information Criterion

PkGu

10/06/2021

Viewpoint from Homo-Frequentistus

In Akaike's work (1979), he generated a reasonable "metric" from the likelihood principle, which is initially denoted by

$$I(f_1, f_0; \Phi) := \int_X \Phi(\frac{f_1}{f_0}) f_0(x)\ dx = \mathbb{E}_{f_0}[\Phi(\frac{f_1}{f_0})]$$

where f_0 means the "real" PDF and f_1 is a candidate. For $\mathcal{P}$ parametrized by Θ, and assuming that θ_0 stands for the "real" one, we naturally need the I and ∇I take 0 at θ_0 and the Hessian matrix at θ_0 is positive-definite for sensitivity of $I(\theta, \theta_0; \Phi)$ near θ_0, *i.e.* equations:

$$I(\theta_0, \theta_0; \Phi) = 0;$$

$$\partial_i I(\theta, \theta_0; \Phi)|_{\theta=\theta_0} = 0;$$

$$\partial^2_{i,j} I(\theta, \theta_0; \Phi)|_{\theta=\theta_0} = \Phi^{''}(1) \cdot \int_X \left[\left(\frac{\partial_i f_\theta}{f_\theta}\right)\left(\frac{\partial_j f_\theta}{f_\theta}\right)\right]\Bigg|_{\theta=\theta_0} dx;$$

And consider the multi-variate case: Suppose there are N *i.i.d.* variables for which we do the same calculation while expecting that

$$I_N(\theta_0, \theta_0; \Phi) = 0;$$

$$\partial_i I_N(\theta, \theta_0; \Phi)|_{\theta=\theta_0} = 0;$$

$$\partial^2_{i,j} I_N(\theta, \theta_0; \Phi)|_{\theta=\theta_0} = N \cdot \partial^2_{i,j} I(\theta, \theta_0; \Phi)|_{\theta=\theta_0};$$

It suggests the adoption of $\Phi(r) := log\ r$ for the amount of "information-distance". And it coincides with the definition of *KL divergence* or *relative entropy*.

For an imperfect sub-model, which means that the "real" parameter $\theta \notin \Theta_{sub}$, we can still look for an MLE solution $\hat{\theta_{sub}}$ in Θ_{sub}. For "real" parameter θ_0, we define the loss and risk functions:

$$W(\theta_0; \hat{\theta}) := 2 \cdot KL(\theta_0; \hat{\theta});$$

$$R(\theta_0; \hat{\cdot}) := \mathbb{E}_X[W(\theta_0; \hat{\theta}(x))];$$

Here we encounter the main problem: suppose Θ has local dimension L, from merely *i.i.d.* dataset $X = (x_1, ..., x_N)$ and the model itself to select the "best" submodel Θ_k in which the setting of coordinates is $\theta_{k;k+1} = ... = \theta_{k;L} = 0$. We are able to compute:

$$\omega_{k;L} := -2 \cdot \left(\frac{1}{N} \sum_{n=1}^{N} log \left(\frac{f_{\hat{\theta_k}}(x_n)}{f_{\hat{\theta_L}}(x_n)} \right) \right);$$

as an estimate of $W(\theta_L; \theta_k)$, where θ_k minimizes $W(\theta_L; \_)$ in Θ_k and especially $\theta_L = \theta_0$ is the "real" parameter. And $\eta_{k;L} := N \cdot \omega_{k;L}$, by large sample theory we've already known that

$$\eta_{k;L} \xrightarrow{N \to \infty} \chi^2_{L-k};$$

For simulation of $W(\theta_L; \hat{\theta_k})$, assuming some reasonable smoothness of W, we rather compute the order-2 simulator around θ_L:

$$W_2(\theta_L; \hat{\theta_k}) := (\hat{\theta_k}^i - {\theta_L}^i)(\hat{\theta_k}^j - {\theta_L}^j) I_{ij};$$

where I_{ij} is exactly the i, j-th element of Fisher information matrix at θ_L. We regard I as an inner product, by redefining θ_k to be the I-projection of θ_L to Θ_k, we have

$$W_2(\theta_L; \hat{\theta_k}) = \|\hat{\theta_k} - \theta_L\|_I^2 = \|\hat{\theta_k} - \theta_k\|_I^2 + \|\theta_k - \theta_L\|_I^2;$$

By doing order-2 Taylor expansion of log-likelihood function around respectively $\hat{\theta_L}$ and $\hat{\theta_k}$ and substitute the variable θ by θ_k, we get that up to an order-3 term,

$$\begin{aligned}
\eta_{k;L} &= N(\|\hat{\theta_L} - \theta_k\|_I^2 - \|\theta_k - \hat{\theta_k}\|_I^2) \\
&= N[\|\hat{\theta_L} - \theta_L\|_I^2 + \|\theta_L - \theta_k\|_I^2 - (\hat{\theta_L} - \theta_L, \theta_L - \theta_k)_I] - N\|\theta_k - \hat{\theta_k}\|_I^2 \\
&= A + B - C - D;
\end{aligned}$$

And under the assumption of smoothness of I, by Taylor expansion of partial-log-likelihood around respectively $\hat{\theta_L}$ and $\hat{\theta_k}$ and substitute the variable θ by θ_k to order 1:

$$\begin{aligned}
&\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \partial_i \; log \; f_{\theta_k}(x_n) \\
&\stackrel{1}{=} \sqrt{N}({\theta_k}^j - \hat{\theta_k}^j) \left[\frac{1}{N} \sum_{n=1}^{N} \partial_{ij}^2 \; log \; f_{\lambda\hat{\theta_k} + (1-\lambda)\theta_k}(x_n) \right] \\
&\stackrel{2}{=} \sqrt{N}({\theta_k}^j - \hat{\theta_L}^j) \left[\frac{1}{N} \sum_{n=1}^{N} \partial_{ij}^2 \; log \; f_{\lambda\hat{\theta_L} + (1-\lambda)\theta_k}(x_n) \right]
\end{aligned}$$

Noticing that by asymptotic normality, $\sqrt{N}(\hat{\theta_L} - \theta_L)$ have tendency $N(0, I^{-1})$ as $N \to \infty$, we have that the order-1 residue is an asymptotically unbiased estimate of the Fisher information matrix:

$$\frac{1}{N} \sum_{n=1}^{N} \partial_{ij}^2 \; log \; f_{\lambda\hat{\theta_L} + (1-\lambda)\theta_L}(x_n) - I_{ij} = O(\frac{1}{\sqrt{N}})$$

so after identifying $I|_{\theta_L} \approx I|_{\theta_k}$ and taking $I_{ij}{\theta_k}^j = I_{ij}{\theta_L}^j$ into account, we have the equation as $N \to \infty$

$$\sqrt{N}(\hat{\theta_k}^j - \theta_k{}^j)I_{ij} = \sqrt{N}(\hat{\theta_L}^j - \theta_L{}^j)I_{ij}$$

which means that $\hat{\theta_k} - \theta_k$ is almost the I-projection of $\hat{\theta_L} - \theta_L$ onto Θ_k, hence independently and asymptotically $A - D \sim \chi^2_{L-k}$ and $D \sim \chi^2_k$. And B has no relation with N or X while C has expectation 0 and its variation is equal to B. So when N significantly larger than L, B and C are both insignificant compared to $\eta_{k;L}$. So asymptotically

$$N \cdot W_2(\theta_L; \hat{\theta_k}) = \eta_{k;L} - (A - D) + D$$

After taking expectation, we find a good estimator for $W(\theta_L; \hat{\theta_k})$:

$$\hat{W}(\theta_L; \hat{\theta_k}) = N^{-1}(\eta_{k;L} + 2k - L)$$

and we just need to find a best $\hat{\theta_k}$ whose k coordinates minimize $\eta_{k;L} + 2k$.