Model Selection Consistency and BIC

Prof. Dennis Sun

Data 401
Statistical Framework for Evaluating Methods

Given data, we assume a probability model for the data and estimate the parameters in that model.

\[
\theta_0 \xrightarrow{p_{\theta_0}(x)} \text{Data } x \xrightarrow{L(\theta) = p_\theta(x)} \hat{\theta}
\]

How do we know if the estimate is any good or not?

Statisticians assume that there is some true underlying model that generated the data and study how \( \hat{\theta} \) compares to \( \theta_0 \).

Notice that there are 3 “\( \theta \)”s floating around:

- \( \theta_0 \) is the true value of the parameter
- \( \theta \) is the variable that we optimize over to get the MLE
- \( \hat{\theta} \) is the estimated value of the parameter
Two Ways to Evaluate Estimators

• Bias: $E[\hat{\theta}] - \theta_0$.
• Consistency: $\hat{\theta} \to \theta_0$ as $n \to \infty$. 
The MLE is Consistent

Remember that \( \hat{\theta}_n \eqdef \arg \max_{\theta} \log p_\theta(X_1, \ldots, X_n) \).

If \( X_1, \ldots, X_n \) are i.i.d., then:

\[
\hat{\theta}_n = \arg \max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(X_i)
\]

\[
\rightarrow \text{as } n \to \infty
\]

\[
\approx \arg \max_{\theta} E_{\theta_0}[\log p_\theta(X_i)]
\]

\[
= \arg \max_{\theta} \sum_x p_{\theta_0}(x) \log p_\theta(x)
\]

But this is just the negative cross entropy between \( p_{\theta_0} \) and \( p_\theta \). It is optimized when \( p_\theta = p_{\theta_0} \), i.e., when \( \theta = \theta_0 \).

\[= \theta_0.\]

We’ve shown \( \hat{\theta}_n \to \theta_0 \) for the i.i.d. case, but the argument generalizes to the independent, but non-i.i.d. case (i.e., regression).
We would like a model selection procedure that is consistent. That is, if the true model is one of the models being considered, then we want a model selection procedure that chooses the true model as $n \to \infty$.

- For linear regression, leave-one-out cross validation is inconsistent. (Shao 1993)
- But choosing the model with the minimum BIC is consistent. (Schwarz 1978)
Bayesian Information Criterion

The **Bayesian Information Criterion** (BIC) chooses the model with the largest value of

\[
\log p_\theta(Y_1, ..., Y_n) - \frac{d}{2} \log n
\]

where \(d\) is the dimension of the parameter \(\theta\).

**Proof of Consistency in the i.i.d. case:**
Suppose the data are actually generated from some distribution \(q\), which may or may not be in \(p_\theta\):

\[
\frac{1}{n} \sum_{i=1}^{n} \log p_\theta(Y_i) - \frac{1}{n} \frac{d}{2} \log n
\]

\[
\downarrow
\]

\[
e_q[\log p_\theta(Y_i)] - \frac{1}{n} \frac{d}{2} \log n
\]

optimized when \(p_\theta = q\)
Consistency of BIC

\[ E_q[\log p_\theta(Y_i)] - \frac{1}{n} \frac{d}{2} \log n \]

Suppose we are comparing two models. There are two cases to consider:

- **Model 1 contains \( q \), but Model 2 does not.**
  Model 1 will achieve a higher value of \( E_q[\log p_\theta(Y_i)] \) as \( n \rightarrow \infty \), while the second term \( \rightarrow 0 \).

- **Both Model 1 and Model 2 contain \( q \).**
  Both models will achieve the maximum value of \( E_q[\log p_\theta(Y_i)] \), so what matters is the second term. The one with the lower \( d \) wins.

The argument generalizes to the independent, but non-i.i.d. case (i.e., regression).
BIC for Regression

The **Bayesian Information Criterion** (BIC) chooses the model with the largest value of

$$-\log(2\pi \hat{\sigma}^2)^{n/2} - \frac{1}{2\hat{\sigma}^2} \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_d x_{id}))^2 - \frac{d + 2}{2} \log n.$$