

Lecture 9 Bayesian information criterion (BIC)

Recall that under the Laplace approximation,

$$p(\vec{\theta} | \mathcal{D}, M) = \frac{p(\vec{\theta}, \mathcal{D} | M)}{\underbrace{p(\mathcal{D} | M)}_{\text{"Z from before"}}, \text{ where}$$

$$p(\mathcal{D} | M) = e^{\log p(\hat{\vec{\theta}}, \mathcal{D} | M)} (2\pi)^{D/2} |H|^{-1/2}$$

$D = \dim\{\vec{\theta}\}$

Using $\hat{\vec{\theta}} = \vec{\theta}_{\text{MAP}}$, we obtain:

$$\log p(\mathcal{D} | M) = \underbrace{\log p(\mathcal{D} | \vec{\theta}_{\text{MAP}}, M)}_{\text{model type}} + \underbrace{\log p(\vec{\theta}_{\text{MAP}} | M) - \frac{1}{2} \log |H|}_{\text{penalty terms}} + \underbrace{\text{const}}_{\substack{\text{indep. of } M, \\ \text{indep. of } N}}$$

log-likelihood

In the large- N limit, $\vec{\theta}_{\text{MAP}} \rightarrow \vec{\theta}^*$ MLE
and the effects of the prior
are negligible, such that

$$\log p(\mathcal{D} | M) = \log p(\mathcal{D} | \vec{\theta}^*, M) - \underbrace{\frac{1}{2} \log |H|}_{\text{Occam factor}}, \text{ where}$$

$$H_{ij} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \left[\underbrace{-\log p(\mathcal{D} | \vec{\theta}, M) - \log p(\vec{\theta} | M)}_{\substack{\text{discard} \\ \text{the prior}}} \right] \bigg|_{\vec{\theta}^*}$$

$$\textcircled{=} - \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(\mathcal{D} | \vec{\theta}, M) \Big|_{\vec{\theta}^*}$$

$\underbrace{\sum_{k=1}^N \log p(\mathcal{D}_k | \vec{\theta}, M)}_{N \leftarrow \text{sample size}}$, additive over data points

We can write

$$H = \sum_{k=1}^N H_k = N \langle H \rangle, \text{ then}$$

$$\log |H| = \log |N \langle H \rangle| = \log(N^D \langle H \rangle) \textcircled{=}$$

$$\textcircled{=} D \log N + \underbrace{\log \langle H \rangle}_{\substack{\text{if } H \text{ is} \\ \text{full rank}}}$$

discard, indep. of N

Thus, $\underbrace{\log p(\mathcal{D} | M)}_{\text{BIC score}} \approx \log p(\mathcal{D} | \vec{\theta}^*, M) - \underbrace{\frac{D_M}{2} \log N}_{\text{model complexity penalty}}$

$D_M = \# \text{ parameters in model } M$

The issue of rank: suppose that ($D_M = 3$)

$$H_2 = \begin{pmatrix} H_{11} & H_{12} & 0 \\ H_{21} & H_{22} & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} N \langle H_{11} \rangle & N \langle H_{12} \rangle & 0 \\ N \langle H_{21} \rangle & N \langle H_{22} \rangle & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

due to likelihood only

and that $-\log p(\vec{\theta} | M) = \lambda |\vec{\theta}|^2$, then

$$H = H_z + \lambda \mathbb{I} = \begin{pmatrix} N\langle H_{11} \rangle + \lambda & N\langle H_{12} \rangle & 0 \\ N\langle H_{21} \rangle & \underbrace{N\langle H_{22} \rangle + \lambda}_{\text{neglect}} & 0 \\ 0 & 0 & \underbrace{\lambda}_{\text{neglect}} \end{pmatrix}.$$

$$\begin{aligned} \text{Now, } \det H &\approx N^2 \langle H_{11} \rangle \langle H_{22} \rangle \lambda - N^2 \langle H_{12} \rangle \langle H_{21} \rangle \lambda = \\ &= N^2 \lambda [\langle H_{11} \rangle \langle H_{22} \rangle - \langle H_{12} \rangle \langle H_{21} \rangle]. \end{aligned}$$

Thus,

$$\log |H| = \underbrace{2 \log N}_{\substack{\uparrow \\ \text{not 3!}}} + \underbrace{\text{const}(N)}_{\substack{\text{discard} \\ \text{as before}}}$$
