ADVANCED BAYESIAN MODELING

AIC and DIC

Consider possible data densities

$$p(y \mid \theta)$$

where parameter θ is continuous, with dimension k.

Consider the classical *maximum likelihood estimator* $\hat{\theta}_{\mathsf{mle}}$.

The deviance

$$-2 \log p\big(y \mid \hat{\theta}_{\mathsf{mle}}\big)$$

tends to underestimate $-2\,\mathrm{elpd}_{\hat{\theta}_{\mathsf{mle}}}$.

On average, the underestimation is about $2k$ (BDA3, Sec. 7.2) ...

AIC

The **Akaike Information Criterion (AIC)** is

$$\mathrm{AIC} \;=\; -2\,\log p\big(y \mid \hat{\theta}_{\mathsf{mle}}\big) \;+\; 2k$$

Smaller values are preferred.

Can regard

$$\widehat{\mathrm{elpd}}_{\mathrm{AIC}} \;=\; -\frac{1}{2}\,\mathrm{AIC}$$

as an estimate of $\mathrm{elpd}_{\hat{\theta}_{\mathsf{mle}}}$.

Usage: When comparing several different (non-Bayesian) models, the one with smallest AIC is preferred (for predictive purposes).

Models may be nested (one a special case of another) or not nested.

What if there is prior info? What if model is hierarchical?

What if we want to evaluate a Bayesian estimate instead?

DIC

Consider posterior mean (vector) as a point estimate:

$$\hat{\theta}_{\text{Bayes}} \;=\; \mathrm{E}(\theta \mid y)$$

The **Deviance Information Criterion (DIC)** is

$$\mathrm{DIC} \;=\; -2\,\log p\big(y \mid \hat{\theta}_{\text{Bayes}}\big) \;+\; 2p_{\mathrm{DIC}}$$

where p_{DIC} is called the **effective number of parameters**.

p_{DIC} shouldn't necessarily equal k because

- A strong prior effectively resizes the parameter space
- A hierarchical model might "shrink" separate parameters toward a common value (effectively fewer parameters)

Some proposals:

$$p_{\text{DIC}} = 2\left(\log p\big(y \mid \hat{\theta}_{\text{Bayes}}\big) - \text{E}_{\text{post}}\big(\log p(y \mid \theta)\big)\right)$$

$$p_{\text{DICalt}} = 2\ \text{var}_{\text{post}}\big(\log p(y \mid \theta)\big)$$

where "post" refers to the posterior distribution of θ

$\hat{\theta}_{\text{Bayes}}$, p_{DIC}, and p_{DICalt} can be approximated from a posterior sample $\theta^1, \ldots, \theta^S$.

For example,

$$\text{E}_{\text{post}}\big(\log p(y \mid \theta)\big) \;\approx\; \frac{1}{S}\sum_{s=1}^{S} \log p(y \mid \theta^s)$$

Also, there is a corresponding measure of predictive accuracy

$$\widehat{\text{elpd}}_{\text{DIC}} \;=\; -\frac{1}{2}\,\text{DIC}$$

Usage of DIC is similar to AIC.

Main limitation of DIC: Based on posterior mean, which

- Is just one possible Bayesian point estimate
- Is not invariant to transformation (reparameterization)
- Need not be a good summary of the full posterior
- In extreme cases, may not exist

Changes to Data

Warning:

All models compared using DIC (or AIC) must be for exactly the same data set y.

In particular, different transformations of y (prior to modeling) cannot be directly compared.

Also, if data is reduced or summarized, this must be done for all of the models.

Also, all models must be for exactly the same observations – if observations are dropped from one model (e.g., due to missing data), they must be dropped from all others as well.