ADVANCED BAYESIAN MODELING

Predictive Accuracy

- How well does a given model perform (for prediction)?

- How do two or more models compare? Which is “best”?

Need a way to measure performance ...

Measuring Accuracy

Under a model, possible sampling (data) densities for data y are

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

Consider predicting "new" $\tilde{y}$, which we assume has density

$$p(\tilde{y} \mid \theta)$$

When using a given θ for prediction, there is a cost incurred (or a utility score attained) depending on the actual value of $\tilde{y}$.

Need a function of θ and $\tilde{y}$ to measure this ...

Example:

$$\tilde{y} \quad \sim \quad \mathrm{N}(\theta, 1)$$

A common way to measure cost of using θ as the prediction of $\tilde{y}$ is *squared error loss*:

$$(\tilde{y} - \theta)^2$$

Note: Equivalent to the *logarithmic score*

$$\log p(\tilde{y} \mid \theta) \quad = \quad -\frac{1}{2}\,(\tilde{y} - \theta)^2 \quad + \quad \text{constant}$$

for which larger is better.

Problems:

- Don't know value of $\tilde{y}$ in advance
- True density of $\tilde{y}$ might not be $p(\tilde{y} \mid \theta)$

Thus,

- Use *expected value* of loss (or utility score) over $\tilde{y}$, and
- Take expectation with respect to the "true" density

$$f(\tilde{y})$$

Example (continued): Mean Squared Error of Prediction

$$\mathrm{MSEP}(\theta) \;=\; \mathrm{E}_f(\tilde{y}-\theta)^2 \;=\; \int (\tilde{y}-\theta)^2 f(\tilde{y})\, d\tilde{y}$$

Can be shown that the optimal (minimizing) value of θ is $\mathrm{E}_f(\tilde{y})$

(Often generalized for use in normal-theory regression)

Equivalently, maximize expected score

$$\mathrm{E}_f \log p(\tilde{y} \mid \theta) \;=\; \int \log p(\tilde{y} \mid \theta) f(\tilde{y})\, d\tilde{y}$$

For Point Estimates

Suppose we want to

- Evaluate point estimate $\hat{\theta} \;=\; \hat{\theta}(y)$
- Use logarithmic score $\log p(\tilde{y} \mid \theta)$ to measure predictive accuracy

Then ideally we would use **expected log predictive density**

$$\mathrm{elpd}_{\hat{\theta}} \;=\; \mathrm{E}_f \log p\big(\tilde{y} \mid \hat{\theta}\big)$$

for which larger is better.

Problem: Don't know f.

Deviance

Consider an approximate $\mathrm{elpd}_{\hat{\theta}}$ formed by

- Choosing $\tilde{y}$ to be a replication of data y
- Substituting observed y for $\tilde{y}$, so that the expectation is not needed

After multiplying by -2, the result is the **deviance**

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

for which smaller is better.

Deviance alone is not a good measure of (in)accuracy for a point estimate:

Since point estimate $\hat{\theta}$ is chosen to fit data y, it will appear to have greater accuracy than it really does when evaluated on that same data y.

Hence, the deviance will tend to be smaller than it should be, for estimating $-2\,\mathrm{elpd}_{\hat{\theta}}$.

Need to add a correction factor ...

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:

$$\begin{aligned} p_{\mathrm{DIC}} &= 2\left(\log p(y \mid \hat{\theta}_{\textsf{Bayes}}) - \mathrm{E}_{\textsf{post}}(\log p(y \mid \theta))\right) \\ p_{\mathrm{DIC}\textsf{alt}} &= 2\ \mathrm{var}_{\textsf{post}}(\log p(y \mid \theta)) \end{aligned}$$

where "post" refers to the posterior distribution of θ

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

For example,

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

Also, there is a corresponding measure of predictive accuracy

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

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.

WAIC: A More Bayesian Approach?

Scores Reconsidered

Bayesian methods for predicting $\tilde{y}$ don't use a "plug-in" estimate, as in

$$p\big(\tilde{y} \mid \hat{\theta}\big)$$

but rather use a posterior predictive density

$$p_{\textsf{post}}(\tilde{y}) \;=\; \int p(\tilde{y} \mid \theta)\, p(\theta \mid y)\, d\theta$$

Suggests a different way to define the score for a Bayesian model:

$$\log\, p_{\textsf{post}}(\tilde{y})$$

May seem reasonable to just substitute data y for $\tilde{y}$ and then add a correction, as for DIC (or AIC).

BDA3 recommends something slightly different:

Consider this **expected log pointwise predictive density**

$$\text{elppd} \;=\; \sum_{i=1}^{n} \mathrm{E}_{f_i}\big(\log p_{\mathsf{post}}(\tilde{y}_i)\big)$$

totaled over n separate individual predictions of $\tilde{y}_i$.

Here, f_i is the “true” density of $\tilde{y}_i$.

When $\tilde{y}_i$ is a replication of observation y_i, elppd can be estimated by substitution:

$$\mathrm{lppd} \;\; = \;\; \sum_{i=1}^{n} \log \, p_{\textsf{post}}(y_i)$$

Since the same data y is being used twice (for the posterior and for substitution), this will tend to overestimate elppd.

Needs a correction ...

WAIC

In BDA3, the **Watanabe-Akaike Information Criterion (WAIC)** is

$$\mathrm{WAIC} \;=\; -2\,\mathrm{lppd} \;+\; 2p_{\mathrm{WAIC}}$$

where effective number of parameters p_{WAIC} could be

$$p_{\mathrm{WAIC1}} \;=\; 2\sum_{i=1}^{n}\Big(\log\big(\mathrm{E}_{\textsf{post}}\, p(y_i \mid \theta)\big) \;-\; \mathrm{E}_{\textsf{post}}\big(\log p(y_i \mid \theta)\big)\Big)$$

$$p_{\mathrm{WAIC2}} \;=\; \sum_{i=1}^{n}\mathrm{var}_{\textsf{post}}\big(\log p(y_i \mid \theta)\big)$$

BDA3 recommends second form.

Notes:

- Smaller values of WAIC are preferred (as with AIC or DIC).

- p_{WAIC1} and p_{WAIC2} can be approximated using a posterior Monte Carlo sample: Replace posterior means and variances with sample means and variances.

- WAIC can estimate elppd:

$$\widehat{\mathrm{elppd}}_{\mathrm{WAIC}} \;=\; -\frac{1}{2}\,\mathrm{WAIC} \;=\; \mathrm{lppd} - p_{\mathrm{WAIC}}$$