7: Evaluating, comparing and expanding models

Taylor

University of Virginia

Introduction

This chapter focuses mostly on quantifying a model's predictive capabilities for the purposes of model selection and expansion.

New Notation!

1. f is the true model
2. y is the data we use to estimate our model
3. $\tilde{y}$ is the future (time series) or alternative (not time series) data that we test our predictions on
4. $p_{\text{post}}(\tilde{y}) = p(\tilde{y} \mid y)$
5. $p_{\text{post}}(\theta) = p(\theta \mid y)$
6. $E_{\text{post}}[\cdot]$ is taken with respect to $p(\theta \mid y)$

Definitions

A **scoring rule/function** $S(p, \tilde{y})$ is a function that takes

1. the distribution you're using to forecast p (ppd, or likelihood with estimated parameters), and
2. a realized value $\tilde{y}$

and then gives you a real-valued number/score/utility. Higher is better, although this convention isn't always followed in the literature.

Keep in mind that the realized value cannot be used to fit the data.

Examples

Example: $S(p, \tilde{y}) = -(\tilde{y} - E_p[\tilde{y}])^2$

Example: $S(p, \tilde{y}) = \log p(\tilde{y})$

Definitions

Future/unseen data is unknown, so we must take the expected score under the true distribution f:

$$E_f[S(p, \tilde{y})].$$

A scoring rule is **proper** if the above expectation is minimized when $f = p$.

A scoring rule is **local** if $S(p, \tilde{y})$ only depends on $p(\tilde{y})$ (don't care about events that didn't happen).

Note, when we are dealing with a logarithmic scoring rule, $E[-2 \log p(\tilde{y})]$ is often called an **information criterion.** The book switches back and forth between dealing with expected score, and information criteria.

Examples

Example: $S(p, y) = -(\tilde{y} - E_p[\tilde{y}])^2$
Most common, perhaps not local or proper for non-Gaussian data.

Example: $S(p, y) = \log p(\tilde{y})$
Obviously local. Proper, too (homework question).

Problem

We are generally not able to evaluate the expectation because we don't know f. However, we may be able to wait for new out-of-sample data and use a Monte-Carlo approach:

$$\tilde{n}^{-1} \sum_{i=1}^{\tilde{n}} S(p, \tilde{y}^i) \to E_f[S(p, \tilde{y})]$$

as $\tilde{n} \to \infty$

Problem

We are generally not able to evaluate the expectation because we don't know f. However, we may be able to wait for new out-of-sample data and use a Monte-Carlo approach:

$$\tilde{n}^{-1} \sum_{i=1}^{\tilde{n}} S(p, \tilde{y}^i) \to E_f[S(p, \tilde{y})]$$

as $\tilde{n} \to \infty$

If we can afford to wait for an infinite amount of data, though, what is the point of trying to predict it?

Problem

We are generally not able to evaluate the expectation because we don't know f. However, we may be able to wait for new out-of-sample data and use a Monte-Carlo approach:

$$\tilde{n}^{-1} \sum_{i=1}^{\tilde{n}} S(p, \tilde{y}^i) \to E_f[S(p, \tilde{y})]$$

as $\tilde{n} \to \infty$

If we can afford to wait for an infinite amount of data, though, what is the point of trying to predict it?

NB: textbook looks at the same instead of the average (calls it "elppd").

Another problem

If we're using the ppd, it might not be in closed form. We have to draw $\theta^j \sim p(\theta \mid y)$, too:

$$\tilde{n}^{-1} \sum_{i=1}^{\tilde{n}} \log \left\{ S^{-1} \sum_{j=1}^{S} p(\tilde{y}^i \mid \theta^j) \right\} \to E_f[\log p_{\text{post}}(\tilde{y})]$$

The textbook calls the above quantity multiplied by $\tilde{n}$ the "computed lppd"

A third problem

Don't want to wait for $\tilde{y}$...

and unfortunately, we cannot plug in the same data that we used for estimation. This overestimates the average predictive score.

However, we can get around this in two ways generally:

1. plug in the already-used y data, but then add an extra penalty term (e.g. AIC, DIC, WAIC, etc.)
2. Cross-Validation: split the data y, many different ways, into a train and test set; estimate and evaluate on each split.

Information Criteria

AIC stands for “an information criterion” or “Akaike’s Information Criterion.” Let k be the number of parameters:

$$\widehat{\text{elpd}}_{\text{AIC}} = \log p(y \mid \hat{\theta}_{\text{MLE}}) - \underbrace{k}_{\text{penalty}}$$

or

$$\text{AIC} = \underbrace{-2 \log p(y \mid \hat{\theta}_{\text{MLE}})}_{\text{a deviance}} + 2k$$

We estimate $\hat{\theta}_{\text{MLE}}$ using y, and we plug y into the log likelihood.

Information Criteria

DIC replaces the point estimate with $\hat{\theta}_{\text{Bayes}} = E[\theta \mid y]$, and replaces the penalty term with p_{DIC}

$$\widehat{\text{elpd}}_{\text{DIC}} = \log p(y \mid \hat{\theta}_{\text{Bayes}}) - p_{\text{DIC}}$$

or

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

Information Criteria

The book gives two ways to estimate p_{DIC}:

1. $p_{\text{DIC}} = 2\left(\log p(y \mid \hat{\theta}_{\text{Bayes}} - E_{\text{post}}\left[\log p(y \mid \theta)\right]\right)$
2. $p_{\text{DIC alt}} = 2\,\text{Var}_{\text{post}}\left[\log p(y \mid \theta)\right]$

Both of these can be approximated using samples from the posterior.