Model Assessment and Comparisons

Sudipto Banerjee

1 Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

September 15, 2010

Three critical questions

- Do the inferences from the model make sense?
- Is the model consistent with the data?
- How can we compare and, perhaps, “rank” different plausible models in their order of preference with respect to a given data set?

The role of checking and comparing models

First two stages:

1. Construct a reasonable probability model;
2. Compute the posterior distribution of model parameters — typically by drawing samples from it.

Third stage: Checking the quality of the model’s fit. This is crucial – Prior-to-Posterior inferences involve the whole structure (with hierarchies) of the Bayesian model and can produce spurious inference if the model is poor.

Sensitivity Analysis: How much do posterior inferences change when other probability models are used in place of the present model?

Replicating data sets using the posterior predictive distribution

Let \( y = (y_1, y_2, \ldots, y_n)' \) be the observed data and \( \theta \) be the collection of all parameters (including all hyperparameters) for a model \( p(\theta) \times p(y | \theta) \).

Let \( y_{rep} = (y_{rep,1}, y_{rep,2}, \ldots, y_{rep,n})' \) be the replicated data that we would see if the experiment that produced \( y \) today were replicated with the same model and the same value of \( \theta \) that produced the observed data.

Replicated data \( y_{rep} \) like predictions \( y \), has two components of uncertainty:

- The fundamental variability of the model, represented by the posited variability in the data;
- The posterior uncertainty in the estimation of \( \theta \)

The distribution of \( y_{rep} \) is the posterior predictive distribution:

\[
P( y_{rep} | y ) = \int p( y_{rep} | \theta ) p( \theta | y ) d\theta
\]

We do not evaluate the above integral, but sample from

\[
p( y_{rep} | y ):
\]

- Draw \( \theta^{(i)} \sim p( \theta | y ) \), \( j = 1, 2, \ldots, M \)
- Draw \( y_{rep}^{(i)} \sim p( y_{rep} | \theta^{(i)} ) \), \( j = 1, 2, \ldots, M \).

Usually full inferential output for Bayesian inference comprises a table comprising both samples from the posterior distribution of \( \theta \) and the posterior predictive distribution of replicated data sets.

<table>
<thead>
<tr>
<th>Sample</th>
<th>( \theta_1 )</th>
<th>( \theta_2 )</th>
<th>( \ldots )</th>
<th>( \theta_p )</th>
<th>( y_{rep,1} )</th>
<th>( y_{rep,2} )</th>
<th>( \ldots )</th>
<th>( y_{rep,n} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \theta_1^{(1)} )</td>
<td>( \theta_2^{(1)} )</td>
<td>( \ldots )</td>
<td>( \theta_p^{(1)} )</td>
<td>( y_{rep,1}^{(1)} )</td>
<td>( y_{rep,2}^{(1)} )</td>
<td>( \ldots )</td>
<td>( y_{rep,n}^{(1)} )</td>
</tr>
<tr>
<td>2</td>
<td>( \theta_1^{(2)} )</td>
<td>( \theta_2^{(2)} )</td>
<td>( \ldots )</td>
<td>( \theta_p^{(2)} )</td>
<td>( y_{rep,1}^{(2)} )</td>
<td>( y_{rep,2}^{(2)} )</td>
<td>( \ldots )</td>
<td>( y_{rep,n}^{(2)} )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\ldots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\ldots</td>
<td>\vdots</td>
</tr>
<tr>
<td>M</td>
<td>( \theta_1^{(M)} )</td>
<td>( \theta_2^{(M)} )</td>
<td>( \ldots )</td>
<td>( \theta_p^{(M)} )</td>
<td>( y_{rep,1}^{(M)} )</td>
<td>( y_{rep,2}^{(M)} )</td>
<td>( \ldots )</td>
<td>( y_{rep,n}^{(M)} )</td>
</tr>
</tbody>
</table>
Example: linear regression model

Recall the Bayesian linear regression model with non-informative priors:
\[ y_i | \mu_i, \sigma^2 \sim N(\mu_i, \sigma^2); \quad i = 1, 2, \ldots, n; \]
\[ \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mathbf{x}_i' \mathbf{\beta}; \quad \mathbf{\beta} = (\beta_0, \beta_1, \ldots, \beta_p); \]
\[ \mathbf{\beta}, \sigma^2 \sim p(\mathbf{\beta}, \sigma^2) = \frac{1}{\sigma^2}. \]

Unknown parameters include the regression parameters and the variance, i.e. \( \mathbf{\theta} = (\mathbf{\beta}, \sigma^2) \).

Obtain posterior samples: \( \mathbf{\theta}^{(j)} = (\mathbf{\beta}^{(j)}, \sigma^{2(j)}), j = 1, \ldots, M. \)

In Bayesian inference, a test statistic \( T(\mathbf{y}; \mathbf{\theta}) \) is evaluated over draws from the posterior distribution of the unknown parameters. We call \( T(\mathbf{y}; \mathbf{\theta}) \) a test measure.

The \( p \)-value is computed using the posterior samples of \( \mathbf{\theta} \) and \( \mathbf{y}_{\text{rep}} \).

For each sampled parameter vector \( \mathbf{\theta}^{(j)} = (\mathbf{\beta}^{(j)}, \sigma^{2(j)}), \) we replicate \( n \) data points:
\[ \mathbf{y}_{\text{rep},j}^{(j)} \sim N(\mathbf{x}_i' \mathbf{\beta}^{(j)}, \sigma^{2(j)}), \quad j = 1, \ldots, M \quad \text{and} \quad i = 1, \ldots, n. \]

\( \mathbf{y}_{\text{rep}}^{(j)} = \left( \mathbf{y}_{\text{rep},1}^{(j)}, \mathbf{y}_{\text{rep},2}^{(j)}, \ldots, \mathbf{y}_{\text{rep},n}^{(j)} \right)' \) is the \( j \)-th sample from the posterior predictive distribution \( p(\mathbf{y}_{\text{rep}} | \mathbf{y}) \).

Remark: The number of posterior samples, \( M \), represents post-convergence (i.e. after burn-in) posterior samples. There is no need to consider pre-convergence samples for drawing the posterior predictive samples.
Bayesian p-value

Bayesian p-values close to 0 or 1 signifies lack of fit of the model with respect to the test measure. On the other hand, values of $p_B$ close to 0.5 indicate very good fit. Estimates of $p_B$ may be sensitive to choice of the test measure.

Unlike $p_C$, we should not interpret $p_B$ with regard to “significance levels” of a test. Instead it should be used as a diagnostic to see if the model adequately fits the data. Bayesian $p$-values are not concerned with “Type-I error” rates. Hence, there is no need to consider adjusting $p_B$ for multiple comparisons (in case we use several test measures).

Model comparisons using replicated data

- Compute the posterior predictive mean and variance for each observation:

$$
\mu_{rep,i} = E[y_{rep,i} | y] = \frac{1}{M} \sum_{j=1}^{M} y_{rep,j}^{(i)}; \quad i = 1, \ldots, n;
$$

$$
\sigma^2_{rep,i} = \text{var}[y_{rep,i} | y] = \frac{1}{M} \sum_{j=1}^{M} (y_{rep,j}^{(i)} - \mu_{rep,i})^2.
$$

- Goodness of fit measure $G$ and expected mean-square predictive error $P$:

$$
G = \sum_{i=1}^{n} (y_i - \mu_{rep,i})^2; \quad P = \sum_{i=1}^{n} \sigma^2_{rep,i}; \quad D = G + P
$$

- $D$ is a model comparison metric (lower values better).

Model comparisons using the DIC

- A general choice for the test measure is the deviance:

$$
T(y; \theta) = D(y; \theta) = -2 \log p(y | \theta).
$$

- A better option for hierarchical models that does not require replicated data (saves computation time):

$$
\bar{D}(y) = E[D(y; \theta) | y] = \frac{1}{M} \sum_{j=1}^{M} D(y; \theta^{(j)});
$$

$$
p_D = D(y) - \bar{D}(y); \quad \text{where} \quad \theta = E[\theta | y] = \frac{1}{M} \sum_{j=1}^{M} \theta^{(j)};
$$

$$
DIC = \bar{D}(y) + p_D = 2\bar{D}(y) - D(y; \bar{\theta}).
$$