Model Assessment and Comparisons

Sudipto Banerjee\(^1\) and Andrew O. Finley\(^2\)

\(^1\) Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

\(^2\) Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.

March 4, 2013
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?
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?
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 $\tilde{y}$, has two components of uncertainty:

1. The fundamental variability of the model, represented by the posited variability in the data;

2. 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 $p(y_{rep} | y)$:

1. Draw $\theta^{(j)} \sim p(\theta | y)$, $j = 1, 2, \ldots, M$
2. Draw $y_{rep}^{(j)} \sim p(y_{rep} | \theta^{(j)})$, $j = 1, 2, \ldots, M$. 
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>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</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:

\[ y_i \mid \mu_i, \sigma^2 \overset{ind}{\sim} N(\mu_i, \sigma^2); \quad i = 1, 2, \ldots, n; \]
\[ \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mathbf{x}_i' \boldsymbol{\beta}; \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p); \]
\[ \beta, \sigma^2 \sim p(\beta, \sigma^2) = \frac{1}{\sigma^2}. \]

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

Obtain posterior samples: \( \theta^{(j)} = \{\beta^{(j)}, \sigma^2(j)\} \), \( j = 1, \ldots, M \).
For each sampled parameter vector \( \theta^{(j)} = \{\beta^{(j)}, \sigma^2(j)\} \), we replicate \( n \) data points:

\[
y^{(j)}_{\text{rep},i} \sim N(x_i'\beta^{(j)}, \sigma^2(j)), \quad j = 1, \ldots, M \quad \text{and} \quad i = 1, \ldots, n.
\]

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

**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.
We distinguish between the replicated data, $y_{rep}$, and the predictive outcomes, $\tilde{y}$.

The variable $\tilde{y}$ is any future observable value of the outcome. For example, in a linear regression model $\tilde{y}$ can have its own set of explanatory variables $\tilde{X}$.

On the other hand, $y_{rep}$ must have the same explanatory variables $X$ as those used in the model for the observed data $y$. In this sense, $y_{rep}$ is similar to “predicting the observed data”.
Lack of fit of the data with respect to the posterior predictive distribution can be measured by the tail-area probability, or \( p \)-value, of a test statistic.

Recall the classical \( p \)-value for a test statistic \( T(y) \):

\[
p_C = P \left( T(y_{rep}) \geq T(y) \mid \theta \right),
\]

where the probability is taken over the distribution of \( y_{rep} \) with \( \theta \) fixed (usually at a value specified by a “null” hypothesis).

In classical statistics, the test statistic \( T(y) \) does not depend upon model parameters.
In Bayesian inference, a test statistic can be a function of the parameters and the data because the test measure is evaluated over draws from the posterior distribution of the unknown parameters. We call $T(y; \theta)$ a test measure.

The $p$-value is computed using the posterior samples of $\theta$ and $y_{rep}$. 

---

Classical and Bayesian p-values
Does our model represent our data adequately? Choose a discrepancy measure or test measure, say

\[
T(y; \theta) = T(y; \beta, \sigma^2) = \sum_{i=1}^{n} \frac{(y_i - E[y_i | \theta])^2}{\text{var}(y_i | \theta)}
\]

\[
= \sum_{i=1}^{n} \frac{(y_i - x_i' \beta)^2}{\sigma^2}
\]

Compute \(T(y, \theta^{(j)})\) and the set of of \(T(y_{rep}^{(j)}, \theta^{(j)})\) and obtain “Bayesian \(p\)-values”:

\[
p_B = P(T(y_{rep}, \theta) > T(y, \theta) | y)
\]

\[
= \frac{1}{M} \sum_{j=1}^{M} 1[T(y_{rep}^{(j)}, \theta^{(j)}) > T(y, \theta^{(j)})].
\]
• 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:

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

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

- 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):

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

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

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