

Model Assessment and Comparisons

Sudipto Banerjee¹ and Andrew O. Finley²

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

1

- 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?

2

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?

3

Replicating data sets using the posterior predictive distribution

- Let $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ be the observed data and θ be the collection of *all* parameters (including all hyperparameters) for a model $p(\theta) \times p(\mathbf{y} | \theta)$.
- Let $\mathbf{y}_{rep} = (y_{rep,1}, y_{rep,2}, \dots, y_{rep,n})'$ be the *replicated data* that we *would* see if the experiment that produced \mathbf{y} today were replicated with the same model and the same value of θ that produced the observed data.
- Replicated data \mathbf{y}_{rep} , like predictions $\tilde{\mathbf{y}}$, 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 θ

4

- The distribution of \mathbf{y}_{rep} is the posterior predictive distribution:

$$p(\mathbf{y}_{rep} | \mathbf{y}) = \int p(\mathbf{y}_{rep} | \theta) p(\theta | \mathbf{y}) d\theta$$

- We do not evaluate the above integral, but sample from $p(\mathbf{y}_{rep} | \mathbf{y})$:
 - 1 Draw $\theta^{(j)} \sim p(\theta | \mathbf{y})$, $j = 1, 2, \dots, M$
 - 2 Draw $\mathbf{y}_{rep}^{(j)} \sim p(\mathbf{y}_{rep} | \theta^{(j)})$, $j = 1, 2, \dots, M$.

5

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

Sample	θ_1	θ_2	...	θ_p	$y_{rep,1}$	$y_{rep,2}$...	$y_{rep,n}$
1	$\theta_1^{(1)}$	$\theta_2^{(1)}$...	$\theta_p^{(1)}$	$y_{rep,1}^{(1)}$	$y_{rep,2}^{(1)}$...	$y_{rep,n}^{(1)}$
2	$\theta_1^{(2)}$	$\theta_2^{(2)}$...	$\theta_p^{(2)}$	$y_{rep,1}^{(2)}$	$y_{rep,2}^{(2)}$...	$y_{rep,n}^{(2)}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
M	$\theta_1^{(M)}$	$\theta_2^{(M)}$...	$\theta_p^{(M)}$	$y_{rep,1}^{(M)}$	$y_{rep,2}^{(M)}$...	$y_{rep,n}^{(M)}$

6

Example: linear regression model

- Recall the Bayesian linear regression model with non-informative priors:

$$y_i | \mu_i, \sigma^2 \stackrel{\text{ind}}{\sim} N(\mu_i, \sigma^2); \quad i = 1, 2, \dots, n;$$

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} = \mathbf{x}'_i \boldsymbol{\beta}; \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p);$$

$$\boldsymbol{\beta}, \sigma^2 \sim p(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2}.$$

- Unknown parameters include the regression parameters and the variance, i.e. $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2\}$.
- Obtain posterior samples: $\boldsymbol{\theta}^{(j)} = \{\boldsymbol{\beta}^{(j)}, \sigma^{2(j)}\}$, $j = 1, \dots, M$.

- For each sampled parameter vector $\boldsymbol{\theta}^{(j)} = \{\boldsymbol{\beta}^{(j)}, \sigma^{2(j)}\}$, we replicate n data points:

$$y_{rep,i}^{(j)} \sim N(\mathbf{x}'_i \boldsymbol{\beta}^{(j)}, \sigma^{2(j)}), \quad j = 1, \dots, M \quad \text{and} \quad i = 1, \dots, n.$$

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

- 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.

- We distinguish between the replicated data, \mathbf{y}_{rep} , and the predictive outcomes, $\tilde{\mathbf{y}}$.
- The variable $\tilde{\mathbf{y}}$ is any *future* observable value of the outcome. For example, in a linear regression model $\tilde{\mathbf{y}}$ can have its own set of explanatory variables $\tilde{\mathbf{X}}$.
- On the other hand, \mathbf{y}_{rep} *must* have the same explanatory variables \mathbf{X} as those used in the model for the observed data \mathbf{y} . In this sense, \mathbf{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 $T(\mathbf{y})$:

$$p_C = P(T(\mathbf{y}_{rep}) \geq T(\mathbf{y}) | \boldsymbol{\theta}),$$

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

- In classical statistics, the test statistic $T(\mathbf{y})$ 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 $T(\mathbf{y}; \boldsymbol{\theta})$ a *test measure*.
- The p -value is computed using the posterior samples of $\boldsymbol{\theta}$ and \mathbf{y}_{rep} .

- Does our model represent our data adequately? Choose a discrepancy measure or *test measure*, say

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

$$= \sum_{i=1}^n \frac{(y_i - \mathbf{x}'_i \boldsymbol{\beta})^2}{\sigma^2}$$

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

$$p_B = P(T(\mathbf{y}_{rep}, \boldsymbol{\theta}) > T(\mathbf{y}, \boldsymbol{\theta}) | \mathbf{y})$$

$$= \frac{1}{M} \sum_{j=1}^M 1[T(\mathbf{y}_{rep}^{(j)}, \boldsymbol{\theta}^{(j)}) > T(\mathbf{y}, \boldsymbol{\theta}^{(j)})].$$

- 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:

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

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

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

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

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

Model comparisons using the DIC

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

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

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

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

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

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