Model Checking and Comparison

Lin Zhang

Department of Biostatistics
School of Public Health
University of Minnesota

Model checking and comparison

- We have discussed the first two major steps of a Bayesian analysis:
  - Constructing a Bayesian model – specifying likelihood and priors
  - Calculating the posterior – either analytically or computationally
- Now we need to consider the question *if the model is a good fit* or *which model is the best fit.*
- Three related issues to consider:
  - Robustness: Is any of the model assumptions having an undue impact on the results?
  - Assessment: Does the model provide adequate fit to the data?
  - Selection: Which model (or models) should we choose for final presentation?

Sensitivity analysis

- A Bayesian data analysis is conditional on the validity of the entire structure of the model
  - Correctly specified likelihood
  - Reasonableness of the prior
- **Key question:** If I change the likelihood or the prior, will it change my posterior inference or conclusion?
  - *No:* The model is robust with respect to this assumption .
  - *Yes:* Document the sensitivity, think more carefully about it, and perhaps collect more data.
- Examples of sensitivity tests:
  - likelihood sensitivity: e.g. nonnormal errors; case deletion
    – common tests for both Bayesian and frequentist
  - change in prior distribution: e.g. doubling/halving a prior s.d.; increasing/decreasing a prior mean.
    – robustness in a Bayesian analysis

Model assessment

- **"All models are wrong, but some are useful"** – George Box
- Most (if not all) models do NOT correctly reflect ALL aspects of the data generation process.
  - The goal is not to determine if a model is right or wrong
  - The key question is whether or not the models have a good enough fitness to the data
- You may have models that *differ substantially* that result in *similar* scientific conclusions!

Model assessment

- Practically we examine model's predictive accuracy, which relies on the posterior predictive distribution:

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

  - $\mathbf{y}^{rep}$ denotes replicated data (data that could have been observed under the same model and same value of $\boldsymbol{\theta}$).

- If the model fits, replicated data under the model should look like the observed data.

- Monte Carlo methods are often used to compute the measures of model assessment.

Posterior predictive p-value

- Let $T(\mathbf{y}, \boldsymbol{\theta})$ be a test quantity, or discrepancy measure, which is a scalar summary of parameters and data
- Examples of test quantities:
  - Min or max
  - Mean
  - Sum of squared residuals
- Bayesian posterior predictive p-value:

$$\begin{aligned} p_B &= E_{\boldsymbol{\theta}|\mathbf{y}}[P(T(\mathbf{y}^{rep}, \boldsymbol{\theta}) > T(\mathbf{y}, \boldsymbol{\theta}))] \\ &= \int P(T(\mathbf{y}^{rep}, \boldsymbol{\theta}) > T(\mathbf{y}, \boldsymbol{\theta})|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \end{aligned}$$

  $\Rightarrow$ the probability that the test quantities of replicated data are more extreme than that of observed data

Posterior predictive p-value

- Note that $P(T(\mathbf{y}^{rep}, \boldsymbol{\theta}) > T(\mathbf{y}, \boldsymbol{\theta})|\boldsymbol{\theta})$ parallels a classical p-value, i.e. $P(T(\mathbf{y}^{rep}, \boldsymbol{\theta}) > T(\mathbf{y}, \boldsymbol{\theta})|\boldsymbol{\theta}_0)$, but Bayesian p-value integrates out $\boldsymbol{\theta}$ under the model.

- The Bayesian p-values are easier to compute via Monte Carlo methods!

- Draw $\boldsymbol{\theta}^{(g)} \sim p(\boldsymbol{\theta}|\mathbf{y})$ and then $\mathbf{y}^{rep(g)} \sim p(\mathbf{y}^{rep}|\boldsymbol{\theta}^{(g)})$, and we have

$$\begin{aligned} p_B &= \int \underline{P(T(\mathbf{y}^{rep}, \boldsymbol{\theta}) > T(\mathbf{y}, \boldsymbol{\theta})|\boldsymbol{\theta})}\, p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &= \int \underline{\int I_{T(\mathbf{y}^{rep}, \boldsymbol{\theta}) > T(\mathbf{y}, \boldsymbol{\theta})}\, p(\mathbf{y}^{rep}|\boldsymbol{\theta})}\, p(\boldsymbol{\theta}|\mathbf{y})\, \underline{d\mathbf{y}^{rep}}\, d\boldsymbol{\theta} \\ &\approx \frac{1}{G} \sum_{g=1}^{G} I_{T(\mathbf{y}^{rep(g)}, \boldsymbol{\theta}^{(g)}) > T(\mathbf{y}, \boldsymbol{\theta}^{(g)})} \end{aligned}$$

Posterior predictive p-value

- An extreme Bayesian p-value suggest discrepancy between the fitted model and the observed data.
- Different test quantities can target different aspects of the model
  - Sum of squared residuals: overall model fitness
  - Min/Max: tail behavior
- Caution: Bayesian p-value SHOULD NOT be compared across models – not for model choice!

Marginal checks via cross validation

- So far, we have focused on replicated data from the joint posterior predictive distribution.
- An alternative is to evaluate the marginal prediction for each i separately.
- Use cross-validation marginal predictive distributions:

$$p_i = P(y_i^{rep} \leq y_i | \mathbf{y}_{(-i)})$$

  where $\mathbf{y}_{(-i)}$ denotes the vector of all the data except the i^{th} value.
  - Useful for identifying outliers or checking model calibration
  - Marginal predictive p-values close to 0 or 1 suggest overdispersion.
- Other Marginal checks ...

Cross-validation Residuals

- "Leave-one out" (LOO) residual:

$$r_i = y_i - E(y_i|\mathbf{y}_{(i)}).$$

- Again, Monte Carlo methods can be used for easy computation. Draws $\boldsymbol{\theta}^{(g)} \sim p(\boldsymbol{\theta}|\mathbf{y})$, and we have

$$\begin{aligned} E(y_i|\mathbf{y}_{(i)}) &= E_{\boldsymbol{\theta}}[E(y_i|\mathbf{y}_{(i)}, \boldsymbol{\theta})] \\ &= \int E(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{(i)}) d\boldsymbol{\theta} \\ &\approx \int E(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &\approx \frac{1}{G} \sum_{g=1}^{G} E(y_i|\boldsymbol{\theta}^{(g)}) \, . \end{aligned}$$

  - Approximation should be adequate unless the dataset is small and y_i is an extreme outlier
  - Same $\boldsymbol{\theta}^{(g)}$'s may be used for each $i = 1, \ldots, n$.

Conditional predictive ordinate

- Conditional predictive ordinate (CPO):

$$\begin{aligned} p(y_i|\mathbf{y}_{(i)}) &= \int p(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_{(i)})d\boldsymbol{\theta} \\ &\approx \int p(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \end{aligned}$$

- Similarly, the CPO can be computed via Monte Carlo approximation:

$$p(y_i|\mathbf{y}_{(i)}) \approx \frac{1}{G}\sum_{g=1}^{G} p(y_i|\boldsymbol{\theta}^{(g)}).$$

- Low CPOs indicate poor fit by the model.

- $\log(CPO)$, the log pseudo marginal likelihood (LPML) is often used in practice for computational convenience.

Bayesian Model Selection

- Suppose we want to choose between the models

$$M_1: \quad Y = \beta_0 + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$
$$M_2: \quad Y = \beta_0 + X\beta + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

- Two general ways for model selection in Bayesian framework:
  - Based on predictive performance
  - Bayes factor

Predictive Model Selection

- Evaluating the accuracy of predictions offers a general approach to model comparison.
- Two problems:
  - Measures of predictive accuracy
  - Methods for estimating predictive accuracy

Measures of predictive accuracy

- The predictive performance of a model is generally assessed by scoring functions and rules.
- Scoring functions: refer to measures of predictive accuracy for **point prediction**, which reports a single predicted value.
  - Example: Mean squared error, $-\frac{1}{n}\sum_{i=1}^{n}(y_i^* - E(y_i^*|\theta))^2$
  - MSE is easy to interpret but less appropriate for non-normal data
- Scoring rules: refer to measures of predictive accuracy for **probabilistic prediction**, which captures the full uncertainty in prediction.
  - Example: Log predictive density or log-likelihood, $log(p(y^*|\theta))$
  - The log predictive density is natural for fitness evaluation as it is closed related to the Kullback-Leibler information

Evaluating using log predictive density

- Ideally, predictive performance would be evaluated using external validation data y^*

$$lpd = \log(p(y^*|\mathbf{y})) = \int p(y^*|\theta)p(\theta|\mathbf{y})d\theta$$

- With future data unknown, we average over the distribution of future data $\Rightarrow$ **expected log predictive density**

$$elpd = E(\log(p(y^*|\mathbf{y}))) = \int \log(p(y^*|\mathbf{y}))p(y^*)d\mathbf{y}^*$$

We can plug in $p(\mathbf{y}^*)$ estimate from the data, but this will lead to too optimistic estimate of fitness.

- For n data points, we evaluate predictive accuracy by taking one at a time $\Rightarrow$ **expected log pointwise predictive density**

$$elppd = \sum_{i=1}^{n} E(\log(p(y_i^*|\mathbf{y})))$$

Methods for estimating predictive accuracy

- We will discuss three approaches to estimate out-of sample predictive accuracy using available data:
  - Within-sample accuracy
  - Adjusted within-sample accuracy
  - Cross-validation

Within-sample accuracy

- Naïve estimates using the log predictive accuracy for existing data
- Examples:

$$
\begin{aligned}
\textit{deviance} &= -2 \log p(\mathbf{y}|\hat{\theta}) \\
\textit{lppd} &= \sum_{i=1}^{n} \log \int p(y_i|\theta) p(\theta|\mathbf{y}) d\theta \\
&\approx \sum_{i=1}^{n} \log \left(\frac{1}{G} \sum_{g=1}^{G} p(y_i|\theta^{(g)}) \right)
\end{aligned}
$$

  where $\hat{\theta}$ could be either a frequentist or Bayesian point estimate, and $\theta^{(g)}$ are draws from the posterior $p(\theta|\mathbf{y})$.

- Overestimate the predictive accuracy due to fitting and evaluating model with the same data

Adjusted within-sample accuracy

- Adjust the naïve estimates to account for overfitting
- Usually substract a correction for the number of parameters in the model
- Examples: AIC, DIC, WAIC
- Can give reasonable answers but are only being correct at best in expectation.

Akaike Information Criteria (AIC)

- The most common frequentist approach to adjust within-sample accuracy

$$AIC = -2 \log p(\mathbf{y}|\hat{\theta}_{MLE}) + 2k$$

where $\hat{\theta}_{MLE}$ is the MLE and k is the number of parameters.

- Adds a penalty for model complexity to the deviance,

- Does NOT work well for Bayesian models with hierarchical structure or informative priors, in which the effective number of parameters is difficult to determine

Effective number of parameters

- Consider the one-way ANOVA model

$$Y_i|\theta_i \overset{ind}{\sim} N(\theta_i, 1/\tau_i) \text{ and } \theta_i \overset{iid}{\sim} N(\mu, 1/\lambda),\ i = 1, \ldots, k$$

Suppose μ, λ, and τ_i are known. How many parameters are in this model?

  - If $\lambda = \infty$, all $\theta_i = \mu$ and there are 0 free parameters
  - If $\lambda = 0$, the θ_i are unconstrained and there are k free parameters

- In practice, $0 < \lambda < \infty$ so the "effective number of parameters" is somewhere in between!

- The effective number of parameter reflects the true complexity of a model

- Question: How to determine it?

Deviance Information Criteria

- DIC is the Bayesian analog of AIC, defined as
$$DIC = -2 \log p(\mathbf{y}|\hat{\theta}_{Bayes}) + 2p_{DIC}$$
where MLE is replaced by the posterior mean $\hat{\theta}_{Bayes}$, and k is replaced by the effective number of parameters p_{DIC}

- The effective number of parameters in the DIC is defined as
$$\begin{aligned} p_{DIC} &= 2\left(\log p(\mathbf{y}|\hat{\theta}_{Bayes}) - E_\theta[\log p(\mathbf{y}|\theta)]\right) \\ &\approx 2\left(\log p(\mathbf{y}|\hat{\theta}_{Bayes}) - \frac{1}{G}\sum_{g=1}^{G} \log p(\mathbf{y}|\theta^{(g)})\right) \end{aligned}$$
where $\theta^{(g)}$ are MCMC draws from the posterior $p(\theta|\mathbf{y})$.

- For the one-way ANOVA model, $p_{DIC} = \sum_{i=1}^{k} \frac{\tau_i}{\tau_i+\lambda} \Rightarrow$ Clearly $0 \leq p_{DIC} \leq k$ as desired.

- An alternative expression for p_{DIC} is
$$p_{DIC\ alt} = 2var_\theta[\log p(\mathbf{y}|\theta)]$$

Watanabe-Akaike Information Criteria (WAIC)

- The WAIC is a fully Bayesian version of AIC, defined as

$$\begin{aligned} WAIC &= -2lppd + 2p_{WAIC} \\ &\approx -2\sum_{i=1}^{n} \log\left(\frac{1}{G}\sum_{g=1}^{G} p(y_i|\theta^{(g)})\right) + 2p_{WAIC} \end{aligned}$$

- **Two** estimates for the effective number of parameters:

$$p_{WAIC_1} = 2\sum_{i=1}^{n}\left\{\log(E_\theta[p(y_i|\theta)]) - E_\theta[\log p(y_i|\theta)]\right\}$$

$$\text{or, } p_{WAIC_2} = 2\sum_{i=1}^{n} var_\theta[\log p(y_i|\theta)]$$

  - p_{WAIC_2} is recommended as it gives results closer to LOO-CV.

- WAIC is advantageous over AIC and DIC in that it averages over the posterior distribution rather than conditioning on a point estimate.

Leave-one-out Cross Validation

- Procedure:
  - Partition the data into a training set and a holdout set with a single data point in the holdout
  - Calculate log predictive density for the holdout data, $log(p(y_i|\mathbf{y}_{(-i)}))$
  - Repeat for all data points, and obtain

$$lppd_{loo-cv} = \sum_i log(p(y_i|\mathbf{y}_{(-i)}))$$

- Avoid overfitting
- Can be computationally expensive
- Does NOT work well for structured data, e.g. spatial data, time series, etc.

Model Selection using Bayes factors

- The Bayes factor for comparing two models M_1 versus M_2 is

$$BF(M_2; M_1) = \frac{p(\mathbf{y} \mid M_2)}{p(\mathbf{y} \mid M_1)} = \frac{\int p(\mathbf{y}|\theta_2, M_2)\pi(\theta_2|M_2)d\theta_2}{\int p(\mathbf{y}|\theta_1, M_1)\pi(\theta_1|M_1)d\theta_1}.$$

- When there are multiple models for comparison, we can estimate Bayes factors by sampling over model space:
  - To treat the model indicator M as a parameter, and sample it with other parameters using MCMC, producing a stream of samples $\{M^{(g)}\}_{g=1}^{G}$ from $p(M|\mathbf{y})$.
  - A simple estimate of posterior model probability is then

$$\hat{p}(M = j|\mathbf{y}) = \frac{\#\text{ of } M^{(g)} = j}{\text{total }\#\text{ of } M^{(g)}}$$

  - Bayes factor between any two models can be computed as

$$\widehat{BF}_{jj'} = \frac{\hat{p}(M = j|\mathbf{y})/\hat{p}(M = j'|\mathbf{y})}{p(M = j)/p(M = j')}.$$

Model Selection using Bayes factors

- Strengths:
  - Works well when underlying model is discrete
  - Works for non-nested models
- Weakness:
  - Problematic when underlying model is continuous
  - Requires proper priors
  - Sensitive to dimensionality of the problem