Bayesian Model Diagnostics and Checking

Earvin Balderama

Quantitative Ecology Lab
Department of Forestry and Environmental Resources
North Carolina State University

April 12, 2013
MCMC
MCMC

Steps in (Bayesian) Modeling

1. Model Creation
2. Model Checking (Criticism)
3. Model Comparison
Outline

1. Popular Bayesian Diagnostics
2. Predictive Distributions
3. Posterior Predictive Checks
4. Model Comparison
   - Precision
   - Accuracy
   - Extreme Values
5. Conclusion
Modeling

Classical methods

1. Standardized Pearson residuals
2. $p$-values
3. Likelihood ratio
4. MLE

also apply in Bayesian Analysis;

1. Posterior mean of the standardized residuals.
2. Posterior probabilities
3. Bayes factor
4. Posterior mean
Bayes Factor

For determining which model fits the data better, the Bayes factor is commonly used in a hypothesis test.

**Bayes factor**

The odds ratio of the data favoring Model 1 over Model 2 is given by

\[
BF = \frac{f_1(y)}{f_2(y)} = \frac{\int f(y|\theta_1)f(\theta_1)d\theta_1}{\int f(y|\theta_2)f(\theta_2)d\theta_2}
\]  

(1)

- More robust than frequentist hypothesis testing.
- Difficult to compute, although easy to approximate with software.
- Only defined for proper marginal density functions.
Bayes Factor

- Computation is conditional that one of the models is true.
- Because of this, Gelman thinks Bayes factors are irrelevant.
  - Prefers looking at distance measures between data and model.

Many alternatives to choose from! One of which is...
DIC

Like many good measures of model fit and comparison, the Deviance Information Criterion (DIC) includes

1. how well the model fits the data (*goodness of fit*) and
2. the complexity of the model (*effective number of parameters*).

**Deviance Information Criterion**

\[
\text{DIC} = \bar{D} + p_D
\]
Deviance

\[ D = -2 \log f(y|\theta) \]  \hspace{1cm} (2)

Example: Deviance with Poisson likelihood

\[ D = -2 \log \prod_{i} \left[ \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \right] \]

\[ = -2 \sum_{i} \left[ -\mu_i + y_i \log \mu_i - \log(y_i!) \right] \]
DIC

Like many good measures of model fit and comparison, the Deviance Information Criterion (DIC) includes

1. how well the model fits the data (goodness of fit) and
2. the complexity of the model (effective number of parameters).

Deviance Information Criterion

\[ DIC = \bar{D} + p_D \]
DIC

Like many good measures of model fit and comparison, the Deviance Information Criterion (DIC) includes

1. how well the model fits the data (goodness of fit) and
2. the complexity of the model (effective number of parameters).

Deviance Information Criterion

\[ \text{DIC} = \bar{D} + p_D \]  

where

1. \( \bar{D} = \mathbb{E}(D) \) is the posterior mean of deviance.
2. \( p_D = \bar{D} - D(\bar{\theta}) \) is “mean deviance - deviance at means.”
DIC can then be rewritten as

\[
DIC = 2\bar{D} - D(\bar{\theta})
\]

\[
= D(\bar{\theta}) + 2p_D
\]

\[
= -2 \log f(y|\bar{\theta}) + 2p_D
\]

which is a generalization of

\[
AIC = -2 \log f(y|\hat{\theta}_{MLE}) + 2k
\]  \hspace{1cm} (5)

DIC can be used to compare different models as well as different methods. Preferred models have low DIC values.
DIC

- Requires joint posterior distribution to be approximately multivariate normal.
- Doesn’t work well with
  - highly non-linear models
  - mixture models with discrete parameters
  - models with missing data
- If $p_D$ is negative
  - log-likelihood may be non-concave
  - prior may be misspecified
  - posterior mean may not be a good estimator
Prior Predictive Distribution

One way to decide between competing models is to rank them based on how “well” each model does in predicting future observations. In Bayesian analyses, *predictive distributions* are used for this kind of decision.

Before data is observed, what could we use for predictions?

**Marginal likelihood**

\[
    f(y) = \int f(y|\theta)f(\theta) d\theta
\]

The marginal likelihood is what one would expect data to look like after averaging over the *prior* distribution of \( \theta \), so it is called the *prior predictive distribution*. 

Posterior Predictive Distribution

More interestingly, if a set of data $y$ have already been observed, one can predict future (or new or unobserved) $y'$ from the marginal posterior likelihood of $y'$, called the *posterior predictive distribution*.

Marginal posterior likelihood

$$f(y'|y) = \int f(y'|\theta)f(\theta|y)d\theta$$  \hspace{1cm} (7)

This distribution is what one would expect $y'$ to look like after observing $y$ and averaging over the *posterior* distribution of $\theta$ given $y$. 

Posterior Predictive Checks

Equivalently, \( y' \) can be the missing values and treated as additional parameters to be estimated in a Bayesian framework. Specifically, for each month, we randomly give NA values to 10 observed sites to create a test set of \( n = 220 \) observations, \( \{y_i; i = 1, 2, \ldots, n\} \). Then the \( n \) posterior predictive distributions, \( \{P_i; i = 1, 2, \ldots, n\} \), can then be used to determine measures of overall model goodness-of-fit, as well as predictive performance measures of each individual \( y_i \) in the test set.
Posterior Predictive Ordinate

The posterior predictive ordinate (PPO) is the density of the posterior predictive distribution evaluated at an observation \( y_i \). PPO can be used to estimate the probability of observing \( y_i \) in the future if after having already observed \( y \).

\[
PPO_i = f(y_i|y) = \int f(y_i|\theta)f(\theta|y)d\theta \tag{8}
\]

We can estimate the \( i^{th} \) posterior predictive ordinate by

\[
\widehat{PPO}_i = \frac{1}{S} \sum_{s=1}^{S} f(y_i|\theta^{(s)}) \tag{9}
\]
Posterior Predictive Checks

**BUGS code**

---

**Example: Poisson count model**

```r
model{

  #likelihood
  for(i in 1:N){

    y[i] ~ dpois(mu[i])
    mu[i] <- ...

    ppo.term[i] <- exp(-mu[i] + y[i]*log(mu[i]) - logfact(y[i]))

  }

  #priors
  ...

}
```

Conditional Predictive Ordinate

- PPO is good for prediction, but violates the likelihood principle.
- The conditional predictive ordinate (CPO) is based on leave-one-out-cross-validation.
- CPO estimates the probability of observing \( y_i \) in the future if after having already observed \( y_{-i} \).
- Low CPO values suggest possible outliers, high-leverage and influential observations.
Conditional Predictive Ordinate

\[
CPO_i = f(y_i|\mathbf{y}_{-i}) = \cdots = \left[ \int \frac{1}{f(y_i|\theta)} f(\theta|\mathbf{y}) d\theta \right]^{-1} \quad (10)
\]

Therefore, CPO can be estimated by taking the inverse of the posterior mean of the inverse density function value of \(y_i\) (harmonic mean of the likelihood of \(y_i\)). Thus,

\[
\hat{CPO}_i = \left[ \frac{1}{S} \sum_{s=1}^{S} \frac{1}{f(y_i|\theta^{(s)})} \right]^{-1} \quad (11)
\]
Conditional Predictive Ordinate

Proof

\[ CPO_i = f(y_i | y_{-i}) \] 
\[ = \left[ \frac{f(y_{-i})}{f(y)} \right]^{-1} \] 
\[ = \left[ \int \frac{f(y_{-i} | \theta) f(\theta)}{f(y)} d\theta \right]^{-1} \] 
\[ = \left[ \int \frac{1}{f(y_i | \theta)} \frac{f(y | \theta)}{f(y)} d\theta \right]^{-1} \] 
\[ = \left[ \int \frac{1}{f(y_i | \theta)} f(\theta | y) d\theta \right]^{-1} \] 
\[ = \left[ \mathbb{E}_{\theta | y} \left( \frac{1}{f(y_i | \theta)} \right) \right]^{-1} \]
Predictive Ordinates

Estimate of PPO

\[
\hat{PPO}_i = \frac{1}{S} \sum_{s=1}^{S} f(y_i|\theta^{(s)})
\]  

(18)

Estimate of CPO

\[
\hat{CPO}_i = \left[ \frac{1}{S} \sum_{s=1}^{S} \frac{1}{f(y_i|\theta^{(s)})} \right]^{-1}
\]  

(19)
BUGS code

Example: Poisson count model

model{

  #likelihood
  for(i in 1:N){

    y[i] ~ dpois(mu[i])
    mu[i] <- ...

    ppo.term[i] <- exp(-mu[i] + y[i]*log(mu[i]) - logfact(y[i]))
    icpo.term[i] <- 1/ppo.term[i]
  }

  #priors
  ...
}

LPML

- The sum of the log CPO’s and is an estimator for the log marginal likelihood.
- The “best” model amongst competing models have the largest LPML.

Log-pseudo marginal likelihood

\[
LPML = \frac{1}{n} \sum_{i=1}^{n} \log(CPO_i)
\]  

(20)
Hypothesis Testing

A ratio of LPML’s is a surrogate for the Bayes factor.

Another overall measure for model comparison is the *posterior Bayes factor*, which is simply the Bayes factor but using the posterior predictive distributions.
Measures of Predictive Precision

The mean absolute deviation (MAD) is the mean of the absolute values of the deviations between the actual observed value and the median of its respective \( \mathcal{P} \).

**Mean Absolute Deviation**

\[
MAD = \frac{1}{n} \sum_{i=1}^{n} |y_i - \tilde{\mathcal{P}}_i| \tag{21}
\]

Note: You can also use the median absolute deviation (also MAD!) for a more robust statistic.
Measures of Predictive Precision

- MSE is the mean of the squared deviations (errors) between the actual observed value and the mean of its respective $\mathcal{P}$.
- The average standard deviation of the $\mathcal{P}$’s can also be helpful.

Mean Squared Error

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{\mathcal{P}}_i)^2$$  

(22)

Mean Standard Deviation

$$SD = \frac{1}{n} \sum_{i=1}^{n} \sigma_{\mathcal{P}_i}$$  

(23)
Measures of Predictive Accuracy

Coverage is the proportion of test set observations $y_i$ falling inside some interval of their respective posterior predictive distributions $\mathcal{P}_i$.

90% Coverage

\[
C^{(90\%)} = \frac{1}{n} \sum_{i=1}^{n} \left[ I \left( \mathcal{P}_i^{(.05)} < y_i < \mathcal{P}_i^{(.95)} \right) \right]
\]  

(24)

where $I$ is the indicator function and $\mathcal{P}^{(q)}$ is the estimated $q^{th}$ quantile.

This shows how well the model does in creating posterior predictive distributions that actually capture the true value.
Predictive Performance

Combining information from measures of precision (i.e. Coverage) and measures of accuracy (i.e. SD) is important for model comparison.

A high coverage probability can simply be a result of high-variance posterior predictive distributions.
Prediction of Extreme Values

The Brier score is the squared difference between the posterior predictive probability of exceeding a certain value and whether or not the actual observation exceeds that value. This score can be used as a measure of predictive accuracy of extreme values.

The Brier score for a test set observation, given a certain value $c$, can be computed as

Brier Score

\[
Brier_i = \left[ \mathcal{I}(y_i > c) - \mathcal{P}_i(y_i > c) \right]^2
\]

(25)

where $\mathcal{I}$ is the indicator function and $\mathcal{P}(y > c)$ is the posterior predictive probability that $y_i > c$.
Prediction of Extreme Values

For each observation $y_i$ in the test set, the quantile score can be computed as

**Quantile Score**

$$QS_i = 2 \times \left[ \mathcal{I} \left( y_i < \mathcal{P}_i^{(q)} \right) - q \right] \times \left[ \mathcal{P}_i^{(q)} - y_i \right] \tag{26}$$

where $\mathcal{I}$ is the indicator function and $\mathcal{P}^{(q)}$ is the estimated $q^{th}$ quantile.
Prediction of Extreme Values

The average of the quantile scores and the average of the Brier Scores can be used for model comparison. They evaluate how well a model captures extreme values.

- Smaller scores are better.
- Larger Brier scores suggest lack of predictive accuracy.
- Larger quantile scores suggest that the observed value is very far from its estimated quantile value from $\mathcal{P}$. 
Other Posterior Predictive Checks

1. Other Bayesian $p$-value tests.
2. Gelman Chi-Square tests, and other Chi-Square tests.
3. Quantile Ratio
4. Predictive Concordance
5. Bayesian Predictive Information Criterion (BPIC)
6. L-criterion
7. ...and many more...
Summary

1. Separate your research into the three MC’s.
2. List of model checks is not exhaustive!
3. Choose some based on the focus of your research.
4. Statistics only a guide.