STAT 431 — Applied Bayesian Analysis — Course Notes

Posterior Predictive Checks

Fall 2022

So far, we considered only ways to compare different data model/prior combinations to each other.

How can we check whether a data model/prior combination is a good fit to the data?

Frequentist approach: *lack-of-fit test* (such as a *chi-square test* — later)

Usually produces a p-value — smaller indicates more evidence against the data model.

A Bayesian wants to assess the prior and data model together.

Discrepancies

For notation:

$$
\begin{aligned}
\boldsymbol{y} &= \text{the data (vector)} \\
\boldsymbol{\theta} &= \text{the parameter (vector) in data model } \mathcal{M}
\end{aligned}
$$

We choose a numerical function called a **discrepancy**:

$$T(\boldsymbol{y}; \boldsymbol{\theta})$$

We intended it to measure how far observed data $\boldsymbol{y}$ depart from what would be expected under data model $\mathcal{M}$ with parameter value $\boldsymbol{\theta}$. Larger values should indicate greater departures from the data model.

An example of a discrepancy:

$$T(\boldsymbol{y};\boldsymbol{\theta}) \;=\; \sum_{i=1}^{n} \frac{\big(y_i - \mathrm{E}(Y_i \mid \boldsymbol{\theta})\big)^2}{\mathrm{Var}(Y_i \mid \boldsymbol{\theta})}$$

where the observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$ form a numerical vector, and $(Y_1, \ldots, Y_n)$ has its distribution under the data model.

This is larger when the y_is are generally farther from their means than their variances would suggest, under the data model with parameter value $\boldsymbol{\theta}$.

(Note: Similar in form to the classical *chi-square statistic*.)

Frequentist Approach

A frequentist can't use $T(\boldsymbol{y}; \boldsymbol{\theta})$ directly, since it depends on the unknown $\boldsymbol{\theta}$.

Replacing $\boldsymbol{\theta}$ with an estimate $\hat{\boldsymbol{\theta}}$ might yield a reasonable test statistic

$$D(\boldsymbol{y}) \;=\; T(\boldsymbol{y}; \hat{\boldsymbol{\theta}})$$

but its distribution under data model $\mathcal{M}$ might still depend on the unknown $\boldsymbol{\theta}$.

If the distribution is (approximately) known, a frequentist can compute the (approximate) p-value

$$p \;=\; \mathrm{Prob}\big(D(\widetilde{\boldsymbol{Y}}) \geq D(\boldsymbol{y})\big)$$

where $\widetilde{\boldsymbol{Y}}$ is a *replication* of the data under its model.

In the earlier example:

$$D(\boldsymbol{y}) \quad = \quad \sum_{i=1}^{n} \frac{\left(y_i - \mathrm{E}(Y_i \mid \hat{\boldsymbol{\theta}})\right)^2}{\mathrm{Var}(Y_i \mid \hat{\boldsymbol{\theta}})}$$

is the classical (Pearson) chi-square statistic (where $\hat{\boldsymbol{\theta}}$ is usually an MLE).

When the data model is correct, asymptotic theory often suggests

$$D(\boldsymbol{Y}) \mid \boldsymbol{\theta} \quad \dot{\sim} \quad \chi^2_{n-k}$$

if $\boldsymbol{\theta}$ effectively has k elements.

When this approximation is valid, an approximate p-value is the χ^2_{n-k} PDF tail area to the right of $D(\boldsymbol{y})$.

In contrast, a Bayesian wants to

- assess the prior, not just the data model
- average over a distribution of $\boldsymbol{\theta}$ (to avoid substituting an estimate $\hat{\boldsymbol{\theta}}$)
- avoid any asymptotic approximations

Bayesian Approach

Suppose

$$\widetilde{\boldsymbol{Y}} \mid \boldsymbol{\theta} \;\sim\; \mathcal{M}(\boldsymbol{\theta})$$

is conditionally independent of the data.

Note: Averaging its distribution over the posterior gives the *posterior predictive distribution* of the data.

Note: We can generate $\widetilde{\boldsymbol{Y}}$ for a given $\boldsymbol{\theta}$ by simulating from the data model, which is usually easy.

A simulated value $\widetilde{\boldsymbol{y}}$ would be called a **replicate** data set.

Then, instead of a frequentist p-value, a Bayesian could use a **posterior predictive p-value**

$$p_b \;=\; \mathrm{Prob}\Big(T(\widetilde{\boldsymbol{Y}};\boldsymbol{\theta}) \geq T(\boldsymbol{y};\boldsymbol{\theta}) \;\Big|\; \boldsymbol{y}\Big)$$

where the probability is over the joint posterior distribution of $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{Y}}$.

Sufficiently small p_b indicates a problem with the data model and/or prior.

Usually p_b can't be directly computed. Instead, it can be approximated by posterior simulation (e.g., MCMC):

1. For each posterior-generated value $\boldsymbol{\theta}$, generate a replicate data set $\widetilde{\boldsymbol{y}}$ (conditionally) and compute

$$T(\widetilde{\boldsymbol{y}}; \boldsymbol{\theta}) \quad \text{and} \quad T(\boldsymbol{y}; \boldsymbol{\theta})$$

2. Approximate p_b as the fraction of generated pairs $(\boldsymbol{\theta}, \widetilde{\boldsymbol{y}})$ for which

$$T(\widetilde{\boldsymbol{y}}; \boldsymbol{\theta}) \;\; \geq \;\; T(\boldsymbol{y}; \boldsymbol{\theta})$$

Example: Shark Attacks

Recall:

$$
\begin{aligned}
Y_i &= \text{number of shark attacks (worldwide)} \\
X_i &= \text{year (2005–2017)}
\end{aligned}
$$

Our chosen data model and prior:

$$
\begin{aligned}
Y_i \mid \lambda_i &\sim \textit{indep} \ \text{Poisson}(\lambda_i) \\
\ln(\lambda_i) &= \beta_1 + (X_i - \bar{X})\beta_2 \\
\beta_1, \beta_2 &\sim \textit{iid} \ \text{Normal}(0, 100^2)
\end{aligned}
$$

We will investigate fit using the chi-square discrepancy

$$T(\boldsymbol{y}; \beta_1, \beta_2) \;=\; \sum_{i=1}^{n} \frac{(y_i - \lambda_i)^2}{\lambda_i}$$

(Why is this the chi-square discrepancy?)

If there are problems with the Poisson regression or the (vague) normal priors, we expect $T(\boldsymbol{y}; \beta_1, \beta_2)$ to be large relative to its replicate version, leading to small p_b.

```
data {
  xmean <- mean(x)
  for(i in 1:length(x)) {
    xcent[i] <- x[i] - xmean
  }
}
model {
  for(i in 1:length(y)) {
    y[i] ~ dpois(lambda[i])
    log(lambda[i]) <- beta1 + xcent[i] * beta2

    yrep[i] ~ dpois(lambda[i])
  }

  beta1 ~ dnorm(0, 0.0001)
  beta2 ~ dnorm(0, 0.0001)

  chisq <- sum((y - lambda)^2 / lambda)
  chisqrep <- sum((yrep - lambda)^2 / lambda)
  pb.ind <- chisqrep >= chisq
}
```

Notes:

- The data set used here has only years 2005 to 2017 (and not the missing 2018 observation).
- JAGS automatically vectorizes arithmetic operations.

We will also try an alternative version of the Bayesian model that has a badly mis-specified prior ...

R/JAGS Example 5.2:

Checking a Poisson Regression

Remark: For some other choices of T, obtaining an especially *large* value of p_b (near 1) would also indicate a problem with the data model and/or prior.