Bayesian Analysis

Charlie Gibbons
University of California, Berkeley

ARE 210

Fall 2015

Outline

Comparison of frequentism and Bayesianism

Frequentists are very concerned about data sets that they never see.

Frequentists take expectations over a random X for a fixed θ.

Conditionality principle

If an experiment concerning inference about θ is chosen from a collection of possible experiments independently, then any experiment not chosen is irrelevant to the inference.

Bayesians take expectations over a random θ and condition on the data set actually observed.

Keep in mind what we know, however: that the MLE is asymptotically consistent and efficient.

Any good estimator must equal the MLE asymptotically.

Unbiasedness

Bayesians are not concerned with unbiasedness.

While seemingly desirable, unbiasedness may lead to poor models.

Example: Suppose that the heights of a father and his son follow a bivariate normal distribution with correlation ρ. Unconditionally, the heights have the same mean μ. We know that

$$\mathbb{E}[\, Y_s \mid Y_f = y_f] = \mu + \rho(y_f - \mu).$$

What does ρ have to be for this model to be unbiased?

The Bayesian model

The Bayesian model consists of:

- The *prior probability* of θ: $\pi(\theta)$
- The *sampling distribution* of $\boldsymbol{X}$: $f(\boldsymbol{x})$
- The *likelihood* of the observed sample $\boldsymbol{x}$ for a given θ: $f(\boldsymbol{x} \mid \theta)$

These form the *posterior probability*:

$$\pi(\theta \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \theta)\pi(\theta)}{f(\boldsymbol{x})}.$$

On average, the posterior variance of θ will be smaller than the prior variance of θ.

Choosing a prior

There are three non-exclusive flavors of priors:

- Subjective priors
- Conjugate priors
- Non-informative priors

The prior distribution for any situation is not unique. Yet results can be sensitive to this choice.

Conjugate priors

Conjugate prior

A prior distribution $\pi(\theta)$ is said to be *conjugate* for the likelihood $f(\boldsymbol{x} \mid \theta)$ if the posterior $\pi(\theta \mid \boldsymbol{x})$ is of the same parametric family as the prior $\pi(\theta)$.

Conjugate priors have desirable properties:

- Analytically tractable
- Asymptotically equivalent to the MLE

Conjugates for the exponential family

Consider n i.i.d. draws from an *exponential family* probability distribution. This gives a likelihood of:

$$f(\boldsymbol{x};\eta) = \prod_i h(x_i) \exp\left(\eta' \sum_i T(x) - nA(\eta)\right)$$

($T(x)$ is known as a *sufficient statistic*).

All members of the exponential family have conjugate priors that are also part of the exponential family. They have the form:

$$\pi(\eta;\tau,n_0) = H(\tau,n_0)\exp\left(\tau'\eta - n_0 A(\eta)\right)$$

Finding a conjugate prior

The trick to finding the conjugate is to view the likelihood as a function of θ, rather than $\boldsymbol{x}$, and find a distribution where the random variable component has that form.

Do not attempt to normalize until the end.

Example: Binomial distribution

For a list, see the Wikipedia entry for conjugate prior.

Shrinkage estimators

Conjugate priors often lead to shrinkage estimators.

Shrinkage estimator

A Bayesian estimator is a *shrinkage estimator* if it is a convex combination of the estimate based upon the prior distribution and the MLE.

Non-informative priors

You can always choose a prior to give *any* posterior result that is desired.

Suppose that we don't have any reason to suspect that any particular value for θ is more likely than any other.

Such priors are called *uniform* or *non-informative*.

We may or may not be able to implement non-informative priors using a conjugate distribution. If not, we'd have to use simulation to describe the shape of the posterior.

Example: Binomial distribution

Improper priors

Consider $X \sim \mathrm{N}(\mu, \sigma^2)$, with σ^2 known. A non-informative prior for μ would be

$$\pi_U(\mu) \propto 1$$

along the entire real line.

This is not a valid probability distribution—it is *improper*.

Improper priors may give good results, but *you must ensure that the posterior is proper*.

Transformations

In the binomial case, instead of putting a prior on probability θ, suppose that we modeled the odds

$$\rho = \frac{\theta}{1-\theta}$$

or the log-odds, $\log(\rho)$.

These are all one-to-one transformations of θ, so any choice among them is arbitrary (from a likelihood perspective).

But a uniform prior for θ is *not* a uniform prior for ρ or $\log(\rho)$.

Change of variables

Consider a one-to-one function $\phi = h(\theta)$. Then the prior for ϕ is

$$\pi_\phi(\phi) = \pi_\theta(\theta) \left| \frac{dh(\theta)}{d\theta} \right|.$$

A flat prior is uninformative for a specific parameter, but may be informative for other parameters.

Example: The non-informative prior for the binomial distribution parameter $\log(\rho)$ in p space is Beta$(0, 0)$, which is improper.

Jeffreys priors

The *Jeffreys prior* is the same for any one-to-one reparameterization of the model; the Jeffreys prior can be found either directly or by applying the change-of-variables formula to the Jeffreys prior under a one-to-one transformation.

The Jeffreys prior has the property:

$$\pi_J(\theta) \propto I(\theta)^{1/2},$$

where $I(\theta)$ is the Fisher information:

$$I(\theta) = -\mathbb{E}_X \left[\frac{\partial^2 \log(x;\theta)}{\partial \theta^2} \right]$$

These priors work best when a model has a single parameter.

Example: Binomial distribution

Properties of Jeffreys priors

Jeffreys priors are:

- Invariant to transformation, but are not non-informative for particular transformations.
- Non necessarily conjugate, but are *limits* of conjugate distributions.

Consider: What is Jeffreys prior equivalent to in a frequentist world?

Maximizing "non-informativeness"

"Non-informative" really means that the prior has little influence on the posterior.

"Influence" can be measured by the distance between the posterior and the prior.

The "most non-informative" prior is the one that maximizes the distance between the posterior and the prior.

A common measure of the "distance" between distributions is the *Kullback-Leibler divergence*:

$$\int \pi(\theta \mid \boldsymbol{x}) \log \left(\frac{\pi(\theta \mid \boldsymbol{x})}{\pi(\theta)} \right) d\theta.$$

Expected KL divergence

The KL divergence is a function of the posterior, which is a function of the data.

Question: How do we choose the prior before seeing the data if the process for choosing the prior involves the data?
Choose the prior based upon *expected* (with respect to $\boldsymbol{X}$) KL divergence:

$$\begin{aligned}
&\int f(\boldsymbol{x}) \int \pi(\theta \mid \boldsymbol{x}) \log\left(\frac{\pi(\theta \mid \boldsymbol{x})}{\pi(\theta)}\right) d\theta \, d\boldsymbol{x} \\
&\quad = \int \int f(\theta, \boldsymbol{x}) \log\left(\frac{\pi(\theta \mid \boldsymbol{x})}{\pi(\theta)}\right) d\theta \, d\boldsymbol{x}.
\end{aligned}$$

Reference priors

Solving for $\pi(\theta)$ gives the *reference prior*.

In general, this is really hard.

In the case of a univariate θ, the reference prior is the Jeffreys prior.

Choosing among uninformative priors

We've seen priors that are:

- Indifferent to particular values of a parameter, but not transformations.
- Indifferent to transformations, but not particular values.

It is not always possible to find both.

If the likelihood component of the model is really dominating the prior, then the results should be insensitive to the choice among relatively uninformative priors.

Now what?

After finding the posterior distribution of θ, we can consider any feature of this distribution.

We may plot the entire distribution, especially if we use a non-conjugate prior.

Features of note are:

- Mean
- Median
- Mode

For inference, we can calculate the probability that θ lies within a particular region.

Credible intervals

Bayesians ask, "what's the probability that θ lies within a particular region?" This forms a *credible interval*:

$$\Pr_\theta(\theta \in C(\boldsymbol{x}) \mid \boldsymbol{x}) \geq 1 - \alpha.$$

Compare to a confidence interval:

$$\Pr_{\boldsymbol{X}}(\theta \in C(\boldsymbol{X})) \geq 1 - \alpha.$$

Credible intervals do not necessary have frequentist coverage probabilities.

HPD regions

Sometimes credible intervals are defined as

$$C(x) = \{\theta \colon \pi(\theta \mid \boldsymbol{x}) \geq K_\alpha\};$$

this is the *highest posterior density (HPD) region.*

Hypothesis testing

Consider the following hypothesis:

$$H_0 : \theta \in \theta_0$$
$$H_1 : \theta \in \theta_1$$

The optimal choice among these two hypotheses is whichever has the highest posterior probability.

Question: What happens if we have a point null?

Posterior odds

Consider the posterior odds:

$$\frac{\pi(H_1 \mid \boldsymbol{x})}{\pi(H_0 \mid \boldsymbol{x})} = \underbrace{\frac{\Pr(\boldsymbol{x} \mid H_1)}{\Pr(\boldsymbol{x} \mid H_0)}}_{\text{Bayes factor}} \underbrace{\frac{\pi(H_1)}{\pi(H_0)}}_{\text{Prior odds}},$$

where the *marginal likelihood* is defined as

$$\Pr(\boldsymbol{x} \mid H) = \int \Pr(\boldsymbol{x} \mid \theta, H) \Pr(\theta \mid H)\, d\theta.$$

Notice that the posterior odds is a function of the prior, but the Bayes factor is only indirectly a function of the prior.

Bayes factor

Because the Bayes factor only indirectly uses the prior, it is a more impartial measure of the evidence against the null.

Jeffreys provided guidelines for evaluating the Bayes factor:

Value ($\log_{10}$)	Conclusion
Less than 0	Evidence for the null
0.0 – 0.5	Weak evidence against the null
0.5 – 1.0	Substantial evidence against the null
1.0 – 2.0	Strong evidence against the null
More than 2	Decisive evidence against the null

Conclusions

Where are we left in the frequentist *v.* Bayesian debate?