Bayesian inference

Gray Calhoun

November 20th, 2014, version 0.8.1

Basics of Bayesian inference (review)

Suppose we know the data are generated as

- Prior: $p(\theta)$
- Likelihood: $p(X \mid \theta)$

Then after observing the data, we can update the prior distribution using Bayes's rule:

$$p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{p(X)}$$

where

$$p(X) = \int p(X \mid \theta)p(\theta)d\theta$$

- This conditional probability is the *posterior density* of θ
- We can also use this approach even if we don't believe that the prior is correct/meaningful

How can we use the posterior density?

- **Point estimation**: Often the posterior mean is consistent and asymptotically equivalent to the MLE

$$\hat{\theta} = \int p(\theta \mid X) d\theta$$

- **Confidence sets**: if we let a and b be the α and $1-\beta$ quantiles of $p(\theta \mid X)$, the interval $[a, b]$ is often a good $1-\alpha-\beta$ confidence interval.
    - Called a *credible interval* in this context.
    - Does not necessarily have correct coverage, but often does.

Informal argument for asymptotic normality of posterior

- Similar to consistency and asymptotic normality of MLE
- Assume θ is in a $1/\sqrt{n}$-neighborhood of θ_0
- Expand log-likelihood around θ_0 (assuming i.i.d. for now)

$$
\begin{aligned}
&\log p(X \mid \theta) - \log p(X \mid \theta_0) \\
&\qquad = \sum_{i=1}^{T} (\log p(x_i \mid \theta) - \log p(X \mid \theta_0)) \\
&\qquad = \sum_{i=1}^{T} \tfrac{\partial}{\partial \theta} \log p(x_i \mid \theta_0)(\theta - \theta_0) \\
&\qquad\quad + \tfrac{1}{2}(\theta - \theta_0)' \Big(\sum_{i=1}^{T} \tfrac{\partial^2}{\partial \theta^2} \log p(x_i \mid \theta_0) \Big)(\theta - \theta_0) + r
\end{aligned}
$$

(regularity conditions like you've seen in 672 ensure that $r = o_p(1/n)$ uniformly in relevant values of θ)

Informal argument for asymptotic normality of posterior

- This lets us expand the log-posterior around θ_0

$$\begin{aligned}
\log p(\theta \mid X) - \log p(\theta_0 \mid X) \\
&= \log p(X \mid \theta) - \log p(X \mid \theta_0) - \log p(\theta) + \log p(\theta_0) \\
&= \sum_{i=1}^{T} \tfrac{\partial}{\partial \theta} \log p(x_i \mid \theta_0)(\theta - \theta_0) \\
&\quad + \tfrac{1}{2}(\theta - \theta_0)' \Big(\sum_{i=1}^{T} \tfrac{\partial^2}{\partial \theta^2} \log p(x_i \mid \theta_0) \Big)(\theta - \theta_0) \\
&\quad - \log p(\theta) + \log p(\theta_0) + r
\end{aligned}$$

Informal argument for asymptotic normality of posterior

- Scale by $1/n$:

$$
\begin{aligned}
\tfrac{1}{n}(\log p(\theta \mid X) - \log p(\theta_0 \mid X)) \\
&= \tfrac{1}{n}\sum_{i=1}^{T} \tfrac{\partial}{\partial\theta} \log p(x_i \mid \theta_0)(\theta - \theta_0) \\
&\quad + (\theta - \theta_0)'\Big(\tfrac{1}{n}\sum_{i=1}^{T} \tfrac{\partial^2}{\partial\theta^2} \log p(x_i \mid \theta_0)\Big)(\theta - \theta_0) \\
&\quad - \tfrac{1}{n}(\log p(\theta) - \log p(\theta_0) - r) \\
&\to^{p} \tfrac{1}{2}(\theta - \theta_0)'\Big(\operatorname{plim} \tfrac{1}{n}\sum_{i=1}^{T} \tfrac{\partial^2}{\partial\theta^2} \log p(x_i \mid \theta_0)\Big)(\theta - \theta_0)
\end{aligned}
$$

Informal argument for asymptotic normality of posterior

- So in large samples, in a neighborhood of θ_0,

$$\log p(\theta \mid X) \approx \log p(\theta_0 \mid X)) + \frac{1}{2}(\theta-\theta_0)'\left(E\sum_{i=1}^{T}\tfrac{\partial^2}{\partial\theta^2}\log p(x_i \mid \theta_0)\right)(\theta-\theta_0)$$

- If $\theta \mid X \sim N(\theta_0, \Sigma)$, we'd have

$$\log p(\theta \mid X) = constant - \tfrac{1}{2}(\theta-\theta_0)'\Sigma^{-1}(\theta-\theta_0)$$

so we have

$$\Sigma \approx -\left(E\sum_{i=1}^{T}\tfrac{\partial^2}{\partial\theta^2}\log p(x_i \mid \theta_0)\right)^{-1}$$

which is also what we see in MLE

Informal argument for asymptotic normality of posterior

- Informally, in large samples where the MLE is consistent and asymptotically normal, the posterior is consistent and asymptotically normal as well for any reasonable prior.
- *Bernstein-von Mises Theorem* (see van der Vaart, 1998, *Asymptotic Statistics*)
- This "proof" is extremely loose. The real proof isn't difficult, but uses more advanced concepts

One reason to be Bayesian: tight coupling with decision theory

- Point forecast for h-steps ahead:

$$\begin{aligned}\hat{y}_{T+h} &= \mathrm{E}(y_{T+h} \,|\, y_1,\ldots,y_T) \\ &= \int \mathrm{E}(y_{T+h} \,|\, \theta, y_1,\ldots,y_{T+h-1}) p(y_{T+h-1} \,|\, \theta, y_1,\ldots,y_{T+h-2}) \ldots \\ &\quad \ldots p(y_{T+1} \,|\, y_1,\ldots,y_T,\theta) p(\theta \,|\, y_1,\ldots,y_T) d\theta \, dy_{T+1} \ldots dy_{T+h-1}\end{aligned}$$

- Density forecast for h-steps ahead:

$$p_y(y_{T+h} \,|\, y_1,\ldots,y_T)$$

- The same estimator gives you the entire joint distribution of parameters and future observations
- For MLE, we'd need to construct separate models for point and density forecasts, and would need to explicitly handle estimation uncertainty

Some other reasons to be Bayesian

- Computational convenience
  - Maximizing the likelihood function can be difficult for some problems
  - For Bayesian inference, we can evaluate all of the intergrals numerically, which can be done much more easily
  - I'm not sure that I buy this rationale very much…
    - but people who actually have experience using these estimators do!
- Shrinkage
- Nuisance parameters
  - *Potentially* many of the modeling decisions we just worried about can be integrated away through judicious choice of prior
  - *Practically* I haven't seen much research on that
- Consistent with accumulation of information over time

Drawbacks of Bayesian approach

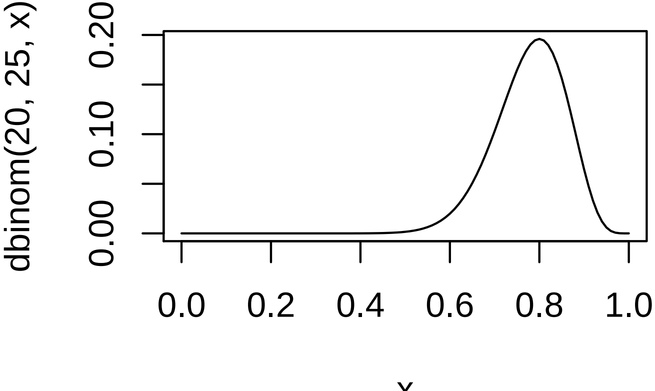
- Some areas are underdeveloped relative to Classical stats
  - HAC covariance matrix adjustment
  - Robustness
  - Nonstationary processes
  - But see recent research by Ulrich Mueller (at Princeton)
- Appropriate priors should be available, just aren't yet
- This (porting robustness, etc. from classical estimators to prior construction) <u>could</u> be an interesting area of research over the next 5 years or so.
  - There's been a lot of recent progress on frequentist theory
  - Talk to me if you're interested in this as a theoretical project
  - There are non-macro areas where the same issues come up (weak identification, potentially)

The simplest example of Bayesian inference you will ever see

- $S \sim binomial(n,p)$, so the likelihood is

$$f_S(s) = \binom{n}{p} p^s (1-p)^{n-s}$$

- Say $n = 25$, $S = 20$, then we can plot the likelihood:
`curve(dbinom(20, 25, x), 0, 1)`



The simplest example of Bayesian inference you will ever see

- $S \sim binomial(n, p)$, so the likelihood is

$$f_S(s) = \binom{n}{s} p^s (1-p)^{n-s}$$

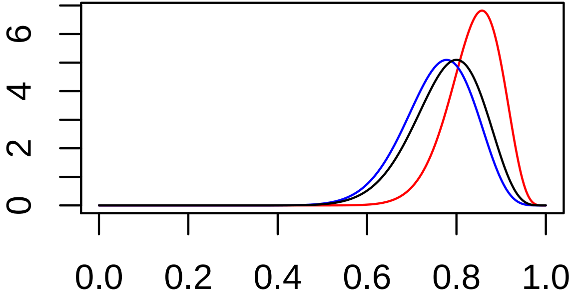
- Now we need a prior density for p. Why not uniform?

$$f_p(p) = 1\{p \in [0,1]\}$$

- Now we can treat likelihood as proportional to posterior density.
- **Conjugate prior** a family of priors is the "conjugate prior" for a family of likelihoods if the posterior density is in the same family.
- $beta(a, b)$ is the conjugate prior for the binomial family and the corresponding posterior is $\beta(a+s, b+n-s)$
  - Prior is "equivalent" to adding a successes and b failures to the dataset
  - $uniform(0,1)$ is the $beta(1,1)$ density
  - Has mean $21/27$ in this example

The simplest example of Bayesian inference you will ever see

Compare posteriors for $beta(1, 1)$ (blue), $beta(0, 0)$ (black), and $beta(10, 0)$ (red) priors



To predict number of successes in next 8 draws

- Prediction is easy. Let S^* be the number of successes in the next 5 draws.
- Use LIE:

$$\begin{aligned}
\Pr[S^* = s \mid S] &= \mathrm{E}(\Pr[S^* = s \mid S, p] \mid S) \\
&= \mathrm{E}(\Pr[S^* = s \mid p] \mid S) \\
&= \mathrm{E}\left(\binom{8}{s} p^s (1-p)^{8-s} \;\middle|\; S \right) \\
&= \binom{8}{s} \int_0^1 p^s (1-p)^{8-s} f_p(p \mid S)
\end{aligned}$$

- Then we (usually) evaluate the probabilities numerically (go to example code)

Key issues to discuss

1. Choosing a prior distribution
2. Working with the posterior numerically
3. If you find this stuff interesting enough that you want to do real research with it, take Stats 544 and (maybe) Stats 644!
    - I will teach you just enough to be dangerous in this class, not enough for you to be confident.
    - Frank Schofheide (UPenn) has several Bayesian Macroeconometrics review articles on his website that look great.

License and copying