Lecture 2 on Bayesian inference

Gray Calhoun

December 2nd, 2014, version 0.9.0

Plan for rest of semester

- Implementing Bayesian estimators: this lecture
  - Typically amounts to simulating from the posterior
  - See Fearnhead (2011) and Chib (2012) review articles
- Prior densities for macro: next lecture
  - Also tie up any remaining loose ends (hah)
- Final exams: next week
  - During classtime as much as possible
  - I will send out a sign-up sheet later today
- First draft of paper: 12/19
- Second draft of paper: 1/30
- Final draft of paper: at your leisure

Quick review

- Parameter of interest: θ
- Prior density: $p(\theta)$
- Likelihood: $p(data \mid \theta)$
- Posterior density:

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

- Obvious next step: how do we use the posterior density once we've calculated it?

Example from last time

- $S \mid \theta \sim \textit{binomial}(n, \theta)$, so the likelihood is

$$p_S(s \mid \theta) = \binom{n}{s} \theta^s (1-\theta)^{n-s}$$

- Uniform prior density:

$$p_\theta(\theta) = 1\{\theta \in [0,1]\}$$

- Posterior:

$$p_\theta(\theta \mid s) = \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \theta^s (1-\theta)^{n-s}$$

which is the $\textit{beta}(s+1, n-s+1)$ density function

Example from last time

- Suppose we want a point estimate of θ
  - Known loss function L ($L(e) = e^2$ is typical)
  - Let's be exotic:

$$L(e) = \begin{cases} 10e^2 & \text{e} < 1/200 \\ e/10 - 1/4000 & e \geq 1/200 \end{cases}$$

  This heavily penalizes large negative errors relative to large positive errors.
- The *risk* (or "Bayes risk") of an estimator $\hat{\theta}$ equals its expected loss:

$$\begin{aligned} \text{risk}(\hat{\theta}) &= \text{E}(L(\theta - \hat{\theta}) \mid data) \\ &= \int L(\theta - \hat{\theta}) p_\theta(\theta \mid data)\, d\theta \end{aligned}$$

  The "Bayes estimator" is the estimator that minimizes the risk:

$$\begin{aligned} \hat{\theta}_B &= \arg\min_{\hat{\theta}} \text{E}(L(\theta - \hat{\theta}) \mid data) \\ &= \arg\min_{\hat{\theta}} \int L(\theta - \hat{\theta}) p_\theta(\theta \mid data)\, d\theta \end{aligned}$$

- Go to R code to finish the example

Need for simulations

- Most quantities we're interested in can be expressed as expectations:

$$\Pr[\theta \leq c \mid data] = E(1\{\theta \leq c\} \mid data)$$

- For simple densities, we can work with the posterior directly
- When there are many parameters, it's better to evaluate the integral through Monte Carlo
  - If $\theta_1, \ldots, \theta_n \sim p(\theta \mid data)$, then (under standard assumptions)

$$\frac{1}{n}\sum_{i=1}^{n} g(\theta_i) \to^p \mathrm{E}(g(\theta) \mid data)$$

as $n \to \infty$

- We can use this to produce the same estimator as before (go to R code)
- Lots of Bayesian inference amounts to simulating data from arbitrary density functions

Rejection sampling

Suppose we can generate data from f, but want to generate data from g. If there is a c s.t. $g(x) \leq cf(x)$ for all x, we can use the Accept-Reject algorithm:

1. Generate a candidate x from f
2. With probability $g(x)/cf(x)$, accept this value of x. Otherwise, go back to step 1.

   Then $x \sim i.i.d.\ g$ (do proof on board)

- Existence of this c is not always guaranteed, and even if it is, finding it can be hard.
- g does not need to be a proper density function, since we can choose c to account for it's missing mass

Importance sampling

Suppose we can generate data from f, but want to generate data from g to calculate the expected value of some $h(X)$ with $X \sim g$.

- Generate $y_1, \ldots, y_n$ from f. Then

$$\mathrm{E}\, h(X) \approx \sum_{i=1}^{n} h(y_i) g(y_i)/f(y_i)$$

- Why?

$$\begin{aligned}
\sum_{i=1}^{n} h(y_i) g(y_i)/f(y_i) &\to^p \mathrm{E}\, h(Y) g(Y)/f(Y) \\
&= \int h(x) \tfrac{g(x)}{f(x)} f(x) dx \\
&= \int h(x) g(x) dx
\end{aligned}$$

where $Y \sim f$

- Doing this naively can be very inefficient.

Markov Chain Monte Carlo

- MCMC is another approach
- A *Markov chain* is a stochastic process $\{x_t\}$ that satisfies

$$p_x(x_t \mid x_{t-1}, x_{t-2}, \ldots) = p_x(x_t \mid x_{t-1})$$

- Stationary Markov chains have three useful properties:
  1. If $X_t \sim p_x$ and $X_{t+1} \mid X_t \sim p_x(\cdot \mid X_t)$, then $X_{t+1} \sim p_x$.
  2. Under weak assumptions, $X_t \mid X_1 \to^d p_x$ as $t \to \infty$
     - Needs to be "irreducible" and "aperiodic"
  3. Markov chains tend to obey the LLN.
- Rather than generate a sequence of independent $\theta_i \sim p_\theta(\cdot \mid data)$, MCMC methods generate $\theta_i \mid \theta_{i-1}$ from a Markov chain

Gibbs sampling

Assume we want to generate draws from the marginal distribution

$$\theta_1, \theta_2, \cdots \sim p_\theta(\cdot \mid data)$$

and we can split each θ_i into several different terms $\theta_{i1}, \theta_{i2}, \ldots, \theta_{ik}$ where

$$\theta_{ij} \mid \theta_{i1}, \ldots, \theta_{i,j-1}, \theta_{i-1,j+1}, \ldots, \theta_{i-1,k} \sim p_j(\cdot \mid \theta_{i1}, \ldots, \theta_{i,j-1}, \theta_{i-1,j+1}, \ldots, \theta_{i-1,k})$$

is easy to simulate from for each j

Gibbs sampling

- The Gibbs sampling algorithm:
  - Start with initial values $\theta_{01}, \theta_{02}, \ldots, \theta_{0k}$
  - For $i = 1, 2, \ldots$ and $j = 1, \ldots, k$, draw

$$\theta_{ij} \mid \theta_{i1}, \ldots, \theta_{i,j-1}, \theta_{i-1,j+1}, \ldots, \theta_{i-1,k} \sim p_j(\cdot \mid \theta_{i1}, \ldots, \theta_{i,j-1}, \theta_{i-1,j+1}, \ldots, \theta_{i-1,k})$$

- Typically people generate and discard the first 1000 or so draws
- This approach works especially well with “auxiliary data”

Quick example: are we in a recession right now?

- Probably not, but we may want to quantify it
- Jim Hamilton has a recession indicator using a simple state-space model (http://econbrowser.com/recession-index)
  - NBER recession dating is slow http://www.nber.org/cycles.html
- I want one too
  - Jim's too responsible to make fun predictions
  - I'd like to see whether we learn much from new data releases
- The model

$$S_t = \begin{cases} 1 & \text{if period } t \text{ is a recession} \\ 2 & \text{if period } t \text{ is not a recession} \end{cases}$$

$$\Pr[S_{t+1} = 1 \mid S_t] = \begin{cases} p & \text{if period } t \text{ is a recession} \\ q & \text{if period } t \text{ is not a recession} \end{cases}$$

$$\Delta\Phi^{-1}(unemployment_t) \mid S_t \sim N(\mu_{S_t}, \sigma^2)$$

- Treat $unemployment_t$ as a known constant

Very simple priors for the unemployment model

- beta(1,1) prior on p and q
- Normal-inverse gamma prior on μ_1, μ_2, σ^2
  - $\mu_1 \sim 1$
  - $\mu_2 \sim 1$
  - $1/\sigma^2 \sim gamma(0,0)$

Likelihood function for unemployment model

- Let $\theta = (p, q, \mu_1, \mu_2, \sigma^2)$
- Let $u_t = \Phi^{-1}(unemployment_t)$
- $u = u_1, \ldots, u_T$
- $S = S_1, \ldots, S_T$
- Likelihood function becomes

$$
\begin{aligned}
f(u \mid \theta) &= \int \cdots \int f(u, S \mid \theta)\, dS_1 \cdots dS_T \\
&= \int \cdots \int \prod_{t=1}^{T} f(u_t, S_t \mid \theta, u_{t-1}, S_{t-1}, \ldots, u_1, S_1)\, dS_1 \cdots dS_T \\
&= \prod_{t=1}^{T} \int f(u_t, S_t \mid \theta, u_{t-1}, S_{t-1}, \ldots, u_1, S_1)\, dS_t \\
&= \prod_{t=1}^{T} \int f(u_t \mid \theta, S_t, u_{t-1}, S_{t-1}, \ldots)\, f(S_t \mid \theta, u_{t-1}, S_{t-1}, \ldots)\, dS_t \\
&= \prod_{t=1}^{T} \int f(u_t \mid S_t, \theta)\, f(S_t \mid S_{t-1}, \theta)\, dS_t
\end{aligned}
$$

Posterior densities, given $S_1, \ldots, S_T$

- Augmenting the dataset lets us use the Gibbs Sampler easily:
  - Generate $\theta \mid u, S$
  - Then generate $S \mid \theta, u$
- Posterior means:
  - $\mu_1 \mid \sigma^2, u, S \sim N(\hat{\mu}_1, \sigma^2/N_1)$
  - $\mu_2 \mid \sigma^2, u, S \sim N(\hat{\mu}_2, \sigma^2/N_2)$

  where
  - $N_i = \sum_{t=1}^{T} 1\{S_t = i\}$
  - $\hat{\mu}_i = (1/N_i) \sum_{t=1}^{T} \Delta\, u_t 1\{S_t = i\}$
- Posterior variance:
  - $1/\sigma^2 \mid u, S \sim gamma(T/2, SSR/2)$

  where
  - $SSR = \sum_{t=1}^{T} (\Delta\, u_t - \mu_{S_t})^2$
- Posterior transition probabilities:
  - $p \mid u, S \sim beta(1 + \hat{P}, 1 + N_1 - \hat{P})$
  - $q \mid u, S \ldots, (u_T, S_T) \sim beta(1 + \hat{Q}, 1 + N_2 - \hat{Q})$

  where
  - $\hat{P} = \sum_{t=1}^{T} 1\{S_t = 1 \text{ and } S_{t-1} = 1\}$
  - $\hat{Q} = \sum_{t=1}^{T} 1\{S_t = 1 \text{ and } S_{t-1} = 2\}$

Simulating $S_1, \ldots, S_T \mid u_1, \ldots, u_T, \theta$

- $S_t \mid \theta, S_{t-1}, u_1, \ldots, u_{t-1}, \theta$ is easy to generate, but does not use the right information set
- We can use Gibbs again to generate each $S_t \mid u, \theta$
  - For $t = 2, \ldots, T-1$, we have

$$\begin{aligned} f_S(s_t \mid \theta, s_1, \ldots, s_{t-1}, s_{t+1}, \ldots, s_T, u) &= f_S(s_t \mid \theta, s_{t-1}, s_{t+1}, u_t) \\ &= \frac{f(s_{t+1}, s_t, u_t \mid \theta, s_{t-1})}{f(s_{t+1}, u_t \mid \theta, s_{t-1})} \\ &\propto f(s_{t+1} \mid s_t, \theta)\, f(u_t \mid s_t, \theta)\, f(s_t \mid s_{t-1}, \theta) \end{aligned}$$

  - For $t = 1$,

$$\begin{aligned} f_S(s_1 \mid \theta, s_2, \ldots, s_T, u) &= f_S(s_1 \mid \theta, s_2, u_1) \\ &\propto f(u_1 \mid s_1, \theta)\, f(s_2 \mid s_1, \theta)\, f(s_1 \mid \theta) \end{aligned}$$

  - For $t = T$,

$$\begin{aligned} f_S(s_T \mid \theta, s_1, \ldots, s_{T-1}, u) &= f_S(s_T \mid \theta, s_{T-1}, u_T) \\ &\propto f(u_T \mid s_T, \theta)\, f(s_T \mid s_{T-1}, \theta) \end{aligned}$$

Putting together the estimator

- Start with an initial guess of $S_{01}, \dots, S_{0T}$
- Repeat the following steps for $i = 1, 2, \dots$
  1. Draw $\theta_{i1} \mid u, S_{i-1,1}, \dots, S_{i-1,T}$
  2. For $t = 1, \dots, T$, draw

$$S_{it} \mid S_{i,t-1}, S_{i-1,t+1}, \theta_i, u$$

- After many iterations, this will generate draws from the correct posterior distribution
- If there's time, we should look at some R code.
- Otherwise, just look at histograms

Metropolis-Hastings

As before, suppose we can draw θ from f, but we want to draw it from g (which we can evaluate)

1. Given a previous draw θ_{i-1}, draw θ_i^* from $f(\cdot;\theta_{i-1})$ (which will typically depend on θ_{i-1})
2. Let $\theta_i = \theta_i^*$ with probability

$$\min\left(\frac{g(\theta^*)\,f(\theta^*;\theta_{i-1})}{g(\theta_{i-1})\,f(\theta_{i-1};\theta^*)}, 1\right).$$

Otherwise let $\theta_i = \theta_{i-1}$

Then $\theta_i, \theta_2, \dots$ forms a Markov Chain and $\theta_t \to^d g$ as $t \to \infty$

- Intuition: similar to rejection sampling: move to regions where the target density is relatively higher.
- For convergence results, etc., see Chib (2012)
- There is an enormous literature on how to implement these samplers well.
- "Random Walk" MH: let $f(\theta^*;\theta_{i-1}) = f(\theta^* - \theta_{i-1})$; often see scaled t-density for f (of course, the scale factor mattters a lot)

Last notes on simulation

- Huge recent literature that we're not touching (even just in macro)
- You know enough to play with these models; please take classes in stats if you want to use them for serious research

Basic prior distributions

- We've already talked about conjugate priors
  - Easy to use
  - Available for some families (binomial, normal, etc)
  - Often one parameterization can be interpreted as "no information"
  - Often unavailable or has other unappealing properties
- "Uninformative" priors
  - "Flat prior" usually isn't uninformative
  - The "Jeffreys prior" is a mostly uninformative prior designed to satisfy some invariance principles
  - "Reference prior" is another (Berger, Bernardo, Sun, 2009)
  - There are even more...
- "Subjective priors"
  - If you actually know something useful about the system you're studying, you can put it into the model as a prior density
  - DSGE models can be used to produce priors
- Empirical Bayes: why not estimate the parameters of the prior?

Priors used in time-series

- First, suppose we have a regression model:

$$y_t = x_t'\beta + e_t$$

where $e_t \mid x_1, \ldots, x_T \sim N(0, \sigma)$

- Conjugate prior for β and σ is Normal-inverse Gamma.
- Start with the priors

$$\beta \mid \sigma \sim N(b, \sigma^2 V)$$
$$1/\sigma^2 \sim gamma(N, \lambda)$$

where b, V, N, and λ are set by the researcher.

Priors used in time-series

- Then we get the posterior

$$\beta \mid \sigma, Y \sim N(b^*, \sigma^2 V^*)$$
$$1/\sigma^2 \mid Y \sim gamma(N+T, \lambda+\lambda^*)$$
$$b^* = V^* V^{-1} b + V^* \sum_{t=1}^{T} x_t y_t$$
$$V^* = (V^{-1} + X'X)^{-1}$$
$$\lambda^* = \sum_{t=1}^{T} (y_t - x_t' \hat{\beta})^2 + (\hat{\beta} - b)' V^{-1} V^* X'X (\hat{\beta} - b)$$

- Interpretation of prior parameters: it's as though we had an additional dataset with

$V^{-1} \approx X'X$ $\quad$ N observations

$b \approx \hat{\beta}$ $\quad$ $\lambda/N \approx \hat{\sigma}^2$

$N, \lambda, V^{-1} \to 0$ is "noninformative"

Priors used for time-series

- Same prior is used for AR(p) and VAR(p)
  - Normal-inverse Gamma is conjugate prior for AR(p) too
  - Normal-inverse Wishart is conjugate prior for VAR(p)
  - Wishart is a multivariate version of the gamma
- "Litterman prior" for a VAR
  - Normal-inverse Wishart
  - Diffuse prior for constant terms
  - For lags of the same variable
    - Coefficient on first lag: $N(1, \gamma^2)$
    - Coefficient on jth lag ($j > 1$): $N(0, (\gamma/j)^2)$
  - For lags of different variables (eq k, variable i)
    - jth lag: $N(0, w\gamma\tau_i/j\tau_k)$
    - has a correction for variances of different series
    - w is a tuning parameter (can be estimated)
  - If series is already differenced (i.e. GDP growth vs. GDP), use 0 for the first lag as well

How do we deal with stationarity more generally?

- Often people don't, or just truncate coefficients to ensure stationarity.
- There are some papers that look at potentially nonstationary priors: Phillips (1991), Berger and Yang (1994), but not many.
- Cointegration is similar: treat the number of cointegrating relationships as known.
- As you can imagine, I find this very unsatisfying.

One brief slide on DSGE models

- One can incorporate DSGE models in (at least) two different ways
  - As the likeihood function: this is analagous to the recession state-space model that we looked at
    - Need to put prior densities on the model's parameters
    - Simulating (well) is more complicated than our simple example
  - As a prior (the likelihood function is then something like a VAR(p))
    1. Generate many draws of the observed variables from the DSGE model
    2. Estimate VAR coefficients and variance on the generated data to get values of b, V, and λ (from the conjugate prior)
    3. Choose N to change the weight that you put on the prior
- In our example: might decide that
  - AR(12) is a good model for Δu_t
  - Shrinking towards the recession state-space model might be good

Future work

This is just a small taste of Bayesian macro. You can read a lot more:

- Fearnhead (2011) and Chib (2012) for more nuance and information on the material discussed in lectures
- Mikusheva's notes for more info on Markov Chains
- Schorfheide's *Bayesian Inference for DSGE Models* http://sites.sas.upenn.edu/schorf/files/dsge_pup_v2_0.pdf
- Geweke's *Complete and Incomplete Econometric Models*

License and copying