Markov Chain Monte Carlo

Charlie Murry

Boston College

April 23, 2019

Overview of Bayesian Estimation

- We wish to know about unkown parameter $\theta^0 \in R^N$
- We have data y, and $L(y|\theta)$ is the likelihood of y given that $\theta = \theta^0$.

Frequentist

- Derive an estimator (MLE) and analyze statistical properties of that estimator

$$\hat{\theta} = \max_{\theta} L(y|\theta).$$

Bayesian

- Start with a prior belief, $p(\theta)$.
- Use data to update their belief to posterior using Bayes Rule

$$\pi(\theta|y) = \frac{L(y|\theta)p(\theta)}{f(y)}$$

where $f(y) = \int L(y|\theta)\pi(\theta)d\theta$ is the marginal distribution of y.

Bayes Rule

$$\pi(\theta|y) = \frac{L(y|\theta)p(\theta)}{f(y)}$$

is just

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)} = \frac{Pr(B|A)Pr(A)}{Pr(B)}.$$

Bayesian Estimation Overview

- Outcome of estimation is $\rho(\cdot)$: summarizes everything we know about where θ is.
- Typically report moments of ρ.
- ρ is typically not tractable. How do we report moments from something that is not tractable.

Good: No need to solve complex optimization problem.

Bad: By construction there is a complex integral.

Tractable Example I

Cameron and Trivedi, p422

Nothing about a posterior per se that requires MCMC.

- Suppose you observe N draws from a normal distribution with mean θ and variance σ^2, e.g.,
$$y_i \sim N(\theta, \sigma^2).$$
- σ^2 is known, but you want to estimate θ.
- A frequentist might use maximum likelihood estimation. The likelihood is:
$$\begin{aligned} L(y|\theta) &= \prod_{i=1}^{N} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{(y_i-\theta)^2}{2\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\sum_{i=1}^{N} \frac{(y_i-\theta)^2}{2\sigma^2}\right\} \\ &\propto \exp\left\{-\frac{N}{2\sigma^2}(\bar{y}-\theta)^2\right\} \end{aligned}$$

Tractable Example II

Cameron and Trivedi, p422

- Clearly this is maximized at $\bar{y}$, which is the MLE estimate.
- Just compute the average from your data.

Bayesian

- Define prior belief, θ.
- Suppose that belief is normally distributed with mean μ and variance τ^2
- Prior density:

$$p(\theta) = (2\pi\tau^2)^{-\frac{1}{2}} \exp\left\{-\frac{(\theta-\mu)^2}{2\tau^2}\right\}.$$

Tractable Example III

Cameron and Trivedi, p422

Following Bayes Rule, the Posterior is proportional to:

$$\begin{aligned}
\pi(\theta|y) &\propto L(y|\theta)p(\theta) \\
&\propto \exp\left\{-\frac{N}{2\sigma^2}(\bar{y}-\theta)^2\right\}\exp\left\{-\frac{(\theta-\mu)^2}{2\tau^2}\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\frac{(\theta-\tilde{\mu})^2}{\tilde{\tau}^2}\right]\right\}
\end{aligned}$$

Where,

$$\tilde{\mu} = \tilde{\tau}^2\left(\frac{N}{\sigma^2}\bar{y} + \frac{1}{\tau^2}\mu\right)$$

$$\tilde{\tau}^2 = \left(\frac{N}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}$$

Tractable Example III

Cameron and Trivedi, p422

- The final line is a normal kernel (just complete the square :)).

- The posterior is normally distributed with mean $\tilde{\mu}$ which is a weighted sum of y and the prior mean μ.

- Since this posterior is normal, it is easy for us to compute the moments.

- Mean of posterior goes to $\bar{y}$ as $N \to \infty$.

- But computing moments of even slightly messier posteriors will require complex integration that we will tackle via simulation.

Review of Monte Carlo Integration

The point of monte carlo integration is to use draws from a distribution to calculate the moments of $\rho(\theta|y)$. If $\rho(\cdot)$ is "easy" to draw from (say, uniform or normal) than we can use traditional monte carlo integration techniques:

$$E[m(\theta)] = \int_{\Theta} m(\theta)\rho(\theta|y)d\theta \approx \frac{1}{S}\sum_{s=1}^{S} m(\theta_s)$$

- $m(\cdot)$ is an arbitrary function, say identity if we want the mean. We need to assume this expectation exists (of course).
- θ_s is a draw from $\rho(\theta|y)$.

However, this isn't helpful if we don't know how to generate draws from $\rho(\cdot)$ and if we did, we could probably just integrate it directly.

Markov Chain Monte Carlo

MCMC uses draws from a Markov Chain, instead of i.i.d. draws from some known distribution.

Use MCMC when:

- Analytic solutions aren't tractable.
- IID sampling doesn't give adequate coverage (perhaps dimension is too high or good approximation of ρ is unknown).

The goal becomes constructing an ergodic Markov Chain F (so that the stationary distribution exists) such that the stationary distribution is exactly ρ. If we do this then we can generate moments of ρ from

$$E[m(\theta)] \approx \frac{1}{S}\sum_{i=1}^{S} m(\theta_i)$$

where $\theta_i \sim F(\cdot|\theta_{i-1})$.

English, please?

ergodic: statistical properties can be deduced from a single, sufficiently long, random sample of the process.

note ergodic: a process that changes erratically at an inconsistent rate

stationary distribution: probability distribution that remains unchanged in the Markov chain as time progresses. (The transition matrix of a discrete processes remains constant)

Markov Chain Theory

Let the state space for θ be discrete, $\Theta = \{\theta^{(1)}, \ldots, \theta^{(K)}\}$.[1]

Let our chain be defined by

$$P(\theta_{r+1} = \theta^{(j)} | \theta_r = \theta^{(i)}) = p_{ij}$$

So the Markov transition matrix is,

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & & & \vdots \\ p_{K1} & p_{K2} & \cdots & p_{KK} \end{bmatrix}$$

[1]Of course this isn't reasonable for estimation but it allows me to skip over a bunch of measure theory.

MC Theory: Stationarity

Let π_0 be an initial distribution over states (a $1 \times K$ vector). Then the distribution over states after 1 period will be:

$$Pr(\theta_1 = \theta^{(j)}) = \sum_{i=1}^{K} Pr(\theta_0 = \theta^{(i)}) p_{ij} = \sum_{i=1}^{K} \pi_{0i} p_{ij}$$

Or in matrix notation for the entire distribution,

$$\pi_1 = \pi_0 P$$

If for all i, j: $p_{ij} > 0$, then every state will be visited infinitely often.

A stationary distribution exists and is unique:

$$\lim_{r \to \infty} \pi_0 P^r = \pi$$

for any π_0 and of course,

$$\pi = \pi P$$

The stationary distribution π is sometimes called the invariant distribution.

Time Reversibility

Definition

A chain is time reversible with respect to π if it has the same behavior backwards and forwards starting from π. That is if the chance of seeing a transition from i to j is the same as seeing a transition from j to i:

$$\pi_i p_{ij} = \pi_j p_{ji}$$

MCMC: Gibbs Sampling

Construct Markov chain by "cycling" through conditional distributions related to π.

- Let $\theta = [\theta_1, \theta_2]'$ with posterior density $p(\theta_1, \theta_2)$.

- *If the conditional densities are known*, then alternating sequential draws from $p(\theta_1 \mid \theta_2)$ and $p(\theta_2 \mid \theta_1)$ converge to $p(\theta_1, \theta_2)$.

Gibbs Example: Probit

Using Data Augmentation

Model:

$$z_i = x_i\beta + \epsilon_i$$

$$y_i = \begin{cases} 0 & z_i \le 0 \\ 1 & z_i > 0 \end{cases}$$

$$\epsilon_i \sim N(0,1)$$

We observe a random sample of (y_i, x_i) and want to estimate β.

Suppose we have a prior $\beta \sim N(\bar{\beta}, A^{-1})$. If we observed z_i then the posterior would be normal (normal is the *conjugate prior* of normal).

However, when z is unobserved there is no simple conjugate prior.

Instead, we can use an "augmentation step" by employing a Gibbs sampler with two blocks (z_i, β), the second step uses draws of z and the normal conjugate prior.

Probit Example — Algorithm

1. Given β_{r-1}, draw z_i by drawing from a truncated normal:

$$z_{i,r}|\beta_{r-1}, y_i, x_i \sim \text{TruncatedNormal}_a^b(-x_i\beta_{r-1}, 1)$$

   Where bounds are $a = 0, b = \infty$ if $y_i = 1$ and $a = -\infty, b = 0$ if $y_i = 0$

2. Draw $\beta_r|z_{i,r}, x_i$ from the posterior of a regression of z on x:

$$\beta_r \sim N(\tilde{\beta}, (X'X + A)^{-1})$$

   where $\tilde{\beta} = (X'X + A)^{-1}(X'z + A\bar{\beta})$.

3. After many draws, we have a sample of β_r which we use as draws from the stationary distribution.

Example — Multivariate Normal

The file `simpleGibbs.m` implements Gibbs sampling to draw from a bivariate normal:

$$(y_1, y_2)' \sim N\left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$$

This trivially implies conditional distributions:

$$y_1|y_2 \sim N(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2), \sigma_1^2(1-\rho^2))$$

Has My MCMC Chain Converged?

Properties of MCMC realy on limit arguments.

The more complicated the chain, the harder it is to make sure you have “converged.”

Some rules of thumb:

- Use a “burn-in” period.
- Plot time series of draws to make sure there is no trend.
- Run chain from several start points.
- Compare distributions from different subsamples of the chain.
- Compute the autocorellation function:

$$s_{\theta_i}(k) = \frac{\sum_{r=k+1}^{R}(\theta_r - \bar{\theta})(\theta_{r-k} - \bar{\theta})}{\sum_{r=1}^{R}(\theta_r - \bar{\theta})^2}$$

  To make sure correlation is dying as time between draws increases.

Has My MCMC Chain Converged? Maybe.

Sadly, there is no proof of convergence in an empirical application.

Similar to finding a global optimizer in frequentist approaches.

MCMC/Bayesian is not a free lunch.

Metropolis Algorithm I

- Gibbs sampling relies on being abble to draw from *conditional* distributions.
- Idea from importance sampling (which we did not go over...)
  - Draw from a known distribution and re-weight.

The Metropolis Algorithm constructs a sequence $\{\theta^{(n)}, n = 1, 2, ...\}$ whose distributions converge to the target posterior.

Metropolis Algorithm II

1. Draw starting point $\theta^{(0)}$ from an initial approximation to the posterior for which $p(\theta^{(0)}) > 0$
   - Ex.: multivariate t-dist centered on mode of marginal posterior distribution.
2. Set $n = 1$. Draw $\theta*$ from a *symmetric* **jumping distribution**, $J_1(\theta^{(1)} \mid \theta^{(0)})$.
   - Ex: $\theta^{(1)} \mid \theta^{(0)} \sim \mathcal{N}(\theta^{(0)}, V)$ for a fixed V.
3. Calculate ratio of densities: $r = p(\theta^*)/p(\theta^{(0)})$.
4. Set
$$\theta^{(1)} = \begin{cases} \theta^* \text{with prob. } \min(r, 1) \\ \theta^{(0)} \text{with prob. } (1 - \min(r, 1)) \end{cases}$$
   so the draw $\theta^{(1)}$ is a draw from a mixture distribution.
5. Return to step 2.
6. Stop after many iterations.

Metropolis Algorithm III

This is *just* a way to increase $p(\theta)$ iteratively.

1. Start with a guess.
2. Update guess if
   2.1 always if likelihood ($p()$) increases or
   2.2 with some probability if likelihood decreases.

Metropolis-Hastings

A refinement that uses a specific **jumping distribution** that is non-symmetric and therefore leads to more frequent transitions.

Go over example

Relationship to “Simulated Annealing”

and other stochastic optimization routines

A non-gradient iterative optimization routine.

1. Perturb one element of parameter vector.
2. Replace if
   - The objective function improves or
   - The objective function worsens but “not by too much” according to a “Metropolis” criteria

$$exp((Q_N(\theta_s^*) - Q_N(\hat{\theta}_s^*))/T_s) > u$$

where s is the iterative time, u is a unit uniform draw, and T_s the *temperature*.

Downhill moves are accepted with a probability that decreases over time.

User chooses step size that governs perturbation and the temperature function.

Also see Chernozhukov and Hong, “An MCMC approach to classical estimation”, *J. of Econometrics*, 2003.