Lecture 9:
Bayesian Estimation: Gibbs Sampling

Big Data and Machine Learning for Applied Economics
Econ 4676

Ignacio Sarmiento-Barbieri

Universidad de los Andes

September 7, 2021

Agenda

1. Recap
    - Bayesian Estimation
    - Direct Sampling
2. Gibbs sampling
3. Review & Next Steps
4. Further Readings

Bayesian Estimation

- *Bayes Theorem*

posterior densidad

prior

$$\pi(\beta|X) = \frac{f(X|\beta)\overbrace{p(\beta)}^{\text{prior}}}{m(X)} \tag{1}$$

$\pi(\beta|X) \propto f(X|\beta)\, p(\beta)$

Interes β

- with $m(X)$ is the marginal distribution of X, i.e.

$$m(X) = \int f(X|\beta)p(\beta)d\beta \tag{2}$$

- Bayes' theorem does not tell us what our beliefs should be, it tells us how they should change after seeing new information.

Frequentist Approach

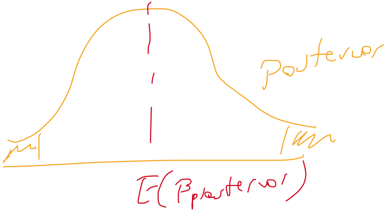
- Suppose the model is

$y = f(x) + u$

$$y_i = \beta x_i + u_i \tag{3}$$
$$u_i \sim N(0, \sigma^2) \tag{4}$$
$$\sigma^2 \text{ is known} \tag{5}$$

- Interest is on some form of $h(\beta)$ e.g. $|\beta|$
- Frequentists
  - $\hat{\beta}_{MLE} = (X'X)^{-1}X'y$ $\quad |\beta|$
  - then use the Delta method (or bootsrap?)
- Bayesians
  - Using simulation based methods $\rightarrow$ direct sampling algorithm

Simulation methods



- Bayesians specify a prior distribution $\beta \sim N(\beta_0, \tau^2)$
- Use conjugate priors and get the posterior

$$\beta | \, Y, X \; \sim \; N\left(\frac{\frac{1}{\sigma^2}\sum_{i=1}^{N} y_i x_i + \frac{1}{\tau^2}\beta_0}{\frac{1}{\sigma^2}\sum_{i=1}^{N} x_i^2 + \frac{1}{\tau^2}}, \frac{1}{\frac{1}{\sigma^2}\sum_{i=1}^{N} x_i^2 + \frac{1}{\tau^2}} \right) \tag{6}$$

- generate i.i.d. samples from the posterior distribution of β, $\pi(\beta|Y)$
- get $h(\beta)$ e.g. $|\beta|$

$E(\beta_{posterior}) = \omega \, \hat{\beta}_{MLE} + (1-\omega)\, \beta_0$

↳ sampling distribution

Direct Sampling

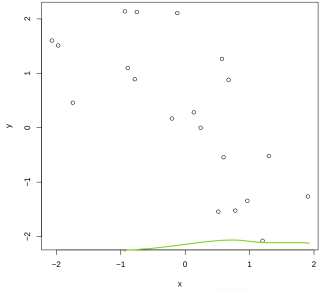
Example: Linear regression

- Goal: posterior mean for $|\beta|$ (quantile $t = 50$)
- I generate data y_i x_i with
  - $y_i = \beta x_i + u_i$, $\quad u_i \sim N(0, \sigma^2)$
  - $N = 20$
  - $\beta_{true} = -1$ *and* $\sigma^2 = 1$
  - $\beta_0 = 0$ *and* $\tau^2 = 100$

posterior ($\pi(\beta|X)$

↳ $|\beta|$

Muestra



$p(\beta) \sim N(0, 100)$

Normal est $N(0,1)$

Direct Sampling

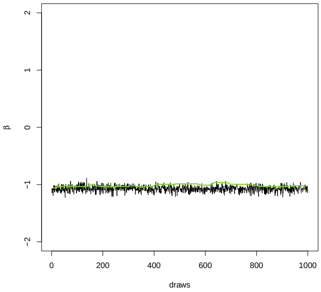
Example: Linear regression

Lik N prior N $\rightarrow$ post $N(m,V)$ 20

- Step 1: we generate S draws from the $N(m, V)$, $\{\beta^s\}$ $1, \dots, S$

$\sigma^2, \Sigma^2, \beta_0, y, X$

Figure 1: Example of draws $\left(\{\beta^s\}_{1,\dots,S}\right)$, $S = 1,000$



β

$\downarrow$

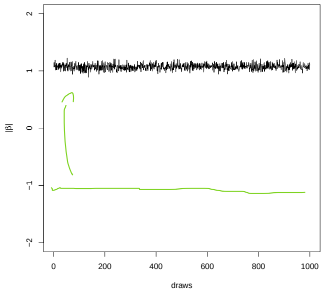
$|\beta|$

Direct Sampling

Example: Linear regression

- Step 2: we are interested in posterior moments of $|\beta|$.
- Turn draws into $\{|\beta|\}_{1,\dots,S}$

Figure 2: Example of draws $\left(\{|\beta^s|\}_{1,\dots,S}\right)$



$y = x_1\beta_1 + x_2\beta_2 + u$

$Cov(\beta_1, \beta_2)$

β^2

$h(\beta) = Cov(\quad)$

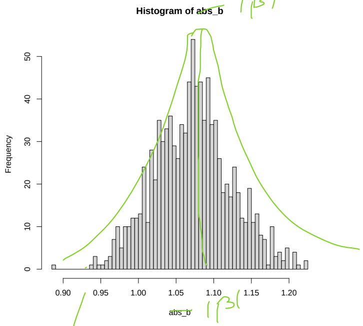
Direct Sampling

Example: Linear regression

$\beta \to \beta$

- Histogram approximation to $\pi\left(|\beta| \;|Y\right)$ using $\{|\beta^s|\}_{1,\ldots,S}$

Figure 3: Example of draws $\left(\{|\beta^s|\}_{1,\ldots,S}\right)$



bandw

Direct Sampling

Example: Linear regression

$\beta \sim N(0, 100)$

y, X

$|\beta| = 1,0719$

- The posterior mean of $|\beta|$ is approximated by

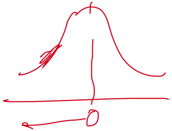
$$E_Y^{\beta}\left[|\beta|\right] \approx \frac{1}{S}\sum_{s=1}^{S}|\beta^s| = 1.0719 \tag{7}$$

- Numerical accuracy use CLT

Gibbs Sampling

- Consider now the linear regression model

$$y_i = \beta x_i + u_i, \quad u_i \sim N(0, \sigma^2)$$



- but assume σ^2 is unknown
- The prior on β is the same as before

$$\beta \; \sim \; N\left(\beta_0, \; \tau^2\right)$$

- We add now a prior on σ^2 is the Inverse-Gamma

$\sigma^2 > 0 \quad \mathbb{R}^+$

$$\sigma^2 \; \sim \; IG\left(a, b\right)$$

- We want to know the joint posterior distribution

$f(x,y) = f(x|y)f(y)$

$$\pi(\beta, \, \sigma^2 \; | Y, X)$$

- Or we want to know the marginal distribution of β and σ^2

Gibbs Sampling

- We know the conditional distribution of β

$$\beta|Y,X,\sigma^2 \sim N(m,V)$$

- We can show that that

$IG(a,b)$

$$\sigma^2|Y,\beta \sim IG\left(a+\frac{N}{2},\ b+\frac{1}{2}\sum_{i=1}^{N}(y_i-x_i\beta)^2\right)$$

- That is, IG is conjugate prior for σ^2

Gibbs Sampling

- Derivation of the conditional posterior distribution

$IG(a, b)$

$$p(\sigma^2 | Y, X, \beta) \propto p\ (Y|X, \sigma^2, \beta)\ p\ (\sigma^2) \tag{8}$$

$$\propto \left(\frac{1}{\sqrt{2\ \pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i - x_i\beta)^2}{\sigma^2}\right) \times \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1}\exp\left(-\frac{b}{\sigma^2}\right)$$

$$\propto (\sigma^2)^{\left(-a\ -1\ -\frac{N}{2}\right)}\exp\left(-\frac{b}{\sigma^2} - \frac{1}{2}\sum_{i=1}^{N}\frac{(y_i - x_i\beta)^2}{\sigma^2}\right)$$

(keep terms that are related to σ^2)

$\sigma^{2\ -(a+\frac{N}{2})-1}\ \exp\left(-\frac{1}{\sigma^2}\left(b + \frac{1}{2}\sum(\quad)^2\right)\right)$

HW

Gibbs Sampling

- Derivation of the conditional posterior distribution
- We can see that the conditional posterior has the IG density form and we can easily deduce that

$$\sigma^2|Y,\beta \sim IG\left(\bar{a}\,\bar{b}\right)$$

- Where

$$\bar{a} = a + \frac{N}{2} \qquad (9)$$
$$\bar{b} = b + \frac{1}{2}\sum_{i=1}^{N}(y_i - x_i\beta)^2$$

Gibbs sampling

- We know conditional posteriors. Can we use these to recover joint distribution?

Gibbs sampling

- We know conditional posteriors. Can we use these to recover joint distribution?
- The answer turns out to be yes
- This is very cool. Joint distribution may be nasty and high-dimensional.
- But, Gibbs sampling allows us to break the nasty joint distribution piece by piece

Gibbs Sampling

Example: Linear regression

- In the context of linear regression example, the Gibbs algorithm works as below
- Enter the following iteration with β^0 *and* $s = 1$

  IG(a, b)

  1. $(\sigma^2)^s \sim p\left(\sigma^2 | Y, \beta^{(s-1)}\right)$
  2. $\beta^s \sim p\left(\beta | Y, \sigma^{(s-1)}\right)$
  3. Go to step 1 with $i = i + 1$ if $i < S$. Otherwise, exit loop

- At the end of the algorithm, you get draws $\left\{\beta^{(s)}, (\sigma^2)^{(s)}\right\}_{s=1,\ldots,S}$.

Gibbs Sampling

Example: Linear regression

- Under regular conditions,

LLN

$$\frac{1}{S - S_0} \sum_{s=S_{0+1}}^{S} h\left(\beta^{(s)}, \left(\sigma^2\right)^{(s)}\right) \rightarrow_{\text{a.s.}} \int h\left(\beta, \sigma^2 | Y\right) d\beta d\sigma^2$$

- Where the first S_0 draws are discarded
- CLT also holds so that we can evaluate the quality of the numerical approximation

GHw

Gibbs Sampling

Example: Linear regression

- Like before, I generate data y_i and x_i with

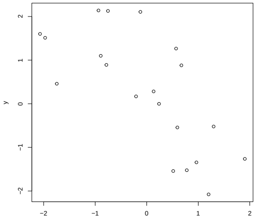
$$y_i = \beta x_i + u_i, \quad u_i \sim N(0, \sigma^2)$$

with

- N = 100
- $\beta_{true} = -1$ and $\sigma^2 = 1$
- Prior for β : $\beta_0 = 0$ and $\tau^2 = 1$
- Prior for σ^2 : $a = 10$, $b = 20$
- Start the algorithm with $\beta^{(0)} = 5$

$N(0,1)$

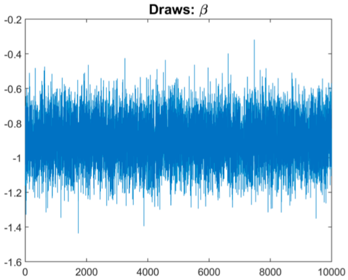
$IG(10,20)$

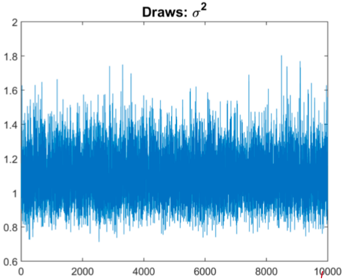


Gibbs Sampling

Example: Linear regression

- Gibbs sampling: Posterior draws

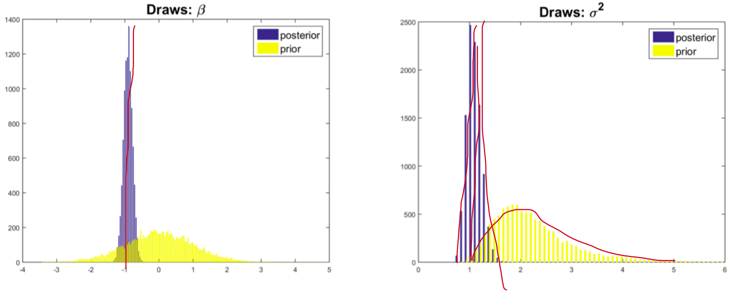




Gibbs Sampling

Example: Linear regression

- Gibbs sampling: Posterior draws



- Once we observe data, we update our belief accordingly.
- Distribution shrinks. Center of the distribution moved toward where the data generated from:

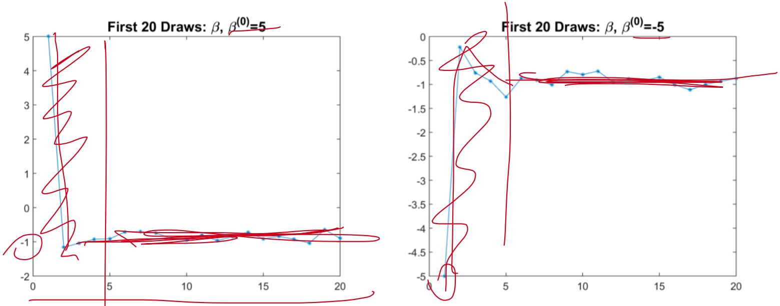
$$\beta_{\text{true}} = -1 \tag{10}$$

$$\sigma^2 = 1 \tag{11}$$

Gibbs Sampling

Initial value effect

- Initial value effect
- We start the algorithm with some arbitrary number β^0. Theory tells us that this initial value should not matter in the long run.



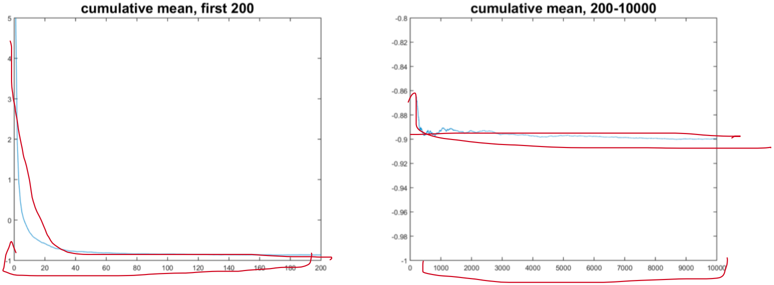
- To get rid of the initial value effect, we usually take out first x draws (say, 1,000 draws)

- Checking convergence
- We start from arbitrary $\beta^{(0)}$
- We can view $\beta^{(0)}$as a draw from the arbitrary distribution (does not have to be the distribution we are interested in).
- Theory tells us that the sequence from Gibbs algorithm converges to the joint posterior distribution.
  - After some iteration, $\beta^{(i)},\ \left(\sigma^2\right)^i$ is a draw from $\pi\left(\beta, \sigma^2|Y\right)$
- How can we check the convergence? There are many ways.

Gibbs Sampling

Checking convergence: Running Mean Plots

- Cumulative mean over draws



- This sequence behaves well in the sense that the Monte Carlo average converges to some number (posterior moment of interest) as the number of draws increases.

Gibbs Sampling

Checking convergence: Geweke's diagnostic check

$y = X\beta + u$

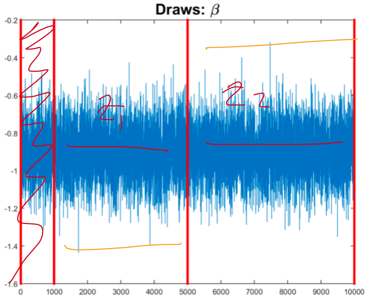
$u \sim N(0, \sigma^2)$

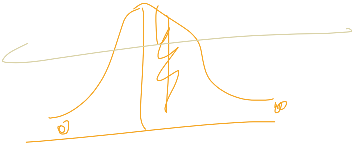
- Take out the first x draws
- Split draws into two non-overlapping parts (e.g. first 10% vs last 50%)
- Compare Z-score

$\beta \sim N(0, \tau^2)$

$\sigma^2 \sim IG(0.1, 0)$

$\frac{x - \mu}{\sigma}$





- Standardized mean (Z-score) of the left sample: -6.9363
- Standardized mean (Z-score) of the right sample: -6.9401
- (you can formally test via hypothesis testing).

$z_1 - z_2$ $\quad y = X\beta +$

Gibbs Sampling

Gibbs sampling versus Direct sampling

β^0, s $\quad \beta^{(1)}$

$\sigma^{2(1)} \sim IG(\bar{a}, \bar{b})$

$\beta^{(1)} \sim N(m, V)$

- A sequence from the Gibbs sampling is serially correlated
  - Previous draw affects the current draw.
  - Gibbs sampler creates a Markov chain. For this reason, it belongs to the class of *Markov chain Monte Carlo (MCMC)* procedures.
- Recall Direct sampling procedure generates i.i.d. draws.

σ_0^2

$\beta \sim N(m, V)$

$\sigma^2 \sim IG(\bar{a}, \bar{b})$

Gibbs Sampling

Gibbs sampling versus Direct sampling

- Both "Direct sampling" and "Gibbs sampling" generate draws that satisfy CLT.
- For draws from direct sampling:

direct sample

$$\sqrt{S}\left(\frac{1}{S}\sum_{i=1}^{N} h\left(\beta^{i}\right)-\int h(\beta)\,\pi(\beta|Y)\,d\beta\right)\rightarrow_{d} N\left(0,V_{\pi}\right)$$

close posterior

- For draws from Gibbs sampling (after discarding first few draws):

Gibbs sample

$$\sqrt{S}\left(\frac{1}{S}\sum_{i=1}^{N} h\left(\beta^{i}\right)-\int h(\beta)\,\pi(\beta|Y)\,d\beta\right)\rightarrow_{d} N\left(0,V_{G}\right)$$

- $V_{\pi} < V_{G}$
- To achieve the same level of approximation error, we need more draws X for Gibbs sampler.

Review

General Gibbs sampler

- Gibbs sampler works for more than two parameters case. Let be θ unknown parameter with $\dim\theta > 1$
- Requirements:
  - Parameter vector θ can be partitioned into $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$
  - For each s it is possible to generate draws of θ_s from the conditional distribution, $p\left(\theta_s|\theta_{s-1}, Y\right)$ where θ_{-s} denotes the vector θ without the partition θ_s
- Gibbs sampler: For $s = 1, \ldots, S$: θ_2 ~ valor inic
  - Draw $\theta_1^{(s+1)}$ from the density $p\left(\theta_1|\theta_2^{(s)}, \theta_3^{(s)}, \ldots, \theta_m^{(s)}, Y\right)$
  - Draw $\theta_2^{(s+1)}$ from the density $p\left(\theta_2|\theta_1^{(s+1)}, \theta_3^{(s)}, \ldots, \theta_m^{(s)}, Y\right)$
  - $\vdots$
  - Draw $\theta_m^{(s+1)}$ from the density $p\left(\theta_m|\theta_1^{(s+1)}, \theta_2^{(s+1)}, \theta_3^{(s+1)}, \ldots, \theta_{m-1}^{(s+1)}, Y\right)$

Review & Next Steps

$IG(a, b)$

$(10, 20)$

STAN

- **Next Class:** Empirical Bayes
- **Next Week:** PS 2

Datos apareados

$y\ x \sim N(0, \sigma^2)$

$\beta \sim N(0, \tau^2)$

$\sigma^2 \sim \chi^2$

Further Readings

- Casella, G., & Berger, R. L. (2002). Statistical inference (Vol. 2, pp. 337-472). Pacific Grove, CA: Duxbury. Chapter 7
- Hoff, P. D. (2009). A first course in Bayesian statistical methods (Vol. 580). New York: Springer.
- Roberts, G.O. and J.S. Rosenthal (2004): General State Space Markov Chains and MCMC algorithms, Probability Surveys, 1, 20–71.
- Geweke, J. (2005): Contemporary Bayesian Econometrics and Statistics, John Wiley & Sons.