

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 density

prior

$$\pi(\beta|X) = \frac{f(X|\beta)p(\beta)}{m(X)} \quad \pi(\beta|X) \propto f(X|\beta)p(\beta) \quad (1)$$

Interes β

\rightarrow 2 a

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

$$m(X) = \int f(X|\beta)p(\beta)d\beta \quad (2)$$

- 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

$$y = f(x) + u$$

$$y_i = \widehat{\beta}x_i + u_i \quad (3)$$

$$u_i \sim N(0, \underline{\sigma^2}) \quad (4)$$

$$\underline{\sigma^2} \text{ is known} \quad (5)$$

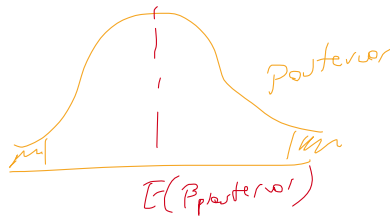
- Interest is on some form of $h(\beta)$ e.g. $|\beta|$
- Frequentists
 - $\hat{\beta}_{MLE} = (X'X)^{-1}X'y$ (p)
 - then use the Delta method (or bootstrap?)

- 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



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

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

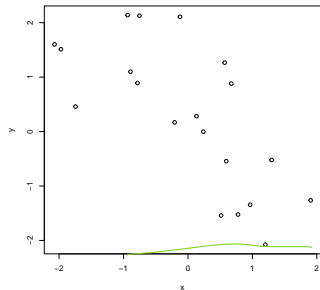
$$E(\beta_{\text{posterior}}) = \omega \hat{\beta}_{MLE} + (1-\omega) \beta_0$$

↳ blending distributions

Direct Sampling

Example: Linear regression

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



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

$p(\beta) \sim N(0, 100)$
kernel est $N(0, 1)$

Direct Sampling

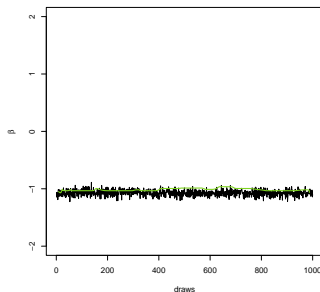
Example: Linear regression

Like N poster $N \rightarrow$ post $N(m, V)$ 20

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

$\sigma^2, \tau^2, \beta_0, y, x$

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



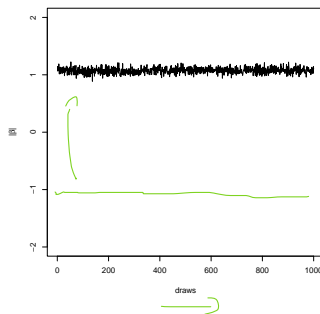
β
 \downarrow
 $|\beta|$

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 $(\{|\beta^s|\}_{1,\dots,S})$



$$y = x_1 \beta_1 + x_2 \beta_2 + u$$
$$\text{COV}(\beta_1, \beta_2)$$

$$\beta^2$$
$$h(\beta) = \omega(\cdot)$$

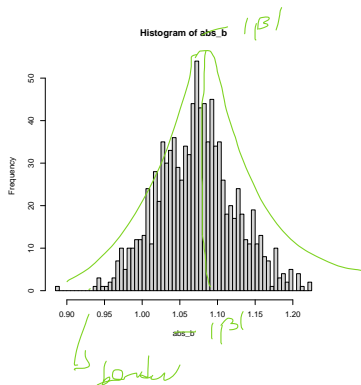
Direct Sampling

Example: Linear regression

$b \rightarrow \hat{b}$

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

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



Direct Sampling

Example: Linear regression

$$p \sim \mathcal{N}(0, 100)$$
$$y, X$$
$$|\beta| = 1.0719$$

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

$$\underline{E_Y^\beta[|\beta|]} \approx \frac{1}{S} \sum_{s=1}^S |\beta^s| = 1.0719 \quad (7)$$

- Numerical accuracy use CLT

Gibbs Sampling

- Consider now the linear regression model

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



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

$$\underline{\beta} \sim \underline{N}(\underline{\beta_0}, \underline{\tau^2})$$

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

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

- We want to know the joint posterior distribution

$$\underline{\pi}(\underline{\beta}, \sigma^2 | Y, X)$$

- Or we want to know the marginal distribution of $\underline{\beta}$ and σ^2

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

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

Gibbs Sampling

- ▶ We know the conditional distribution of β

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

- ▶ We can show that that

$$\text{IG}(a, b)$$

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

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

Gibbs Sampling

► Derivation of the conditional posterior distribution

$$\begin{aligned}
 p(\sigma^2 | Y, X, \beta) &\propto p(Y|X, \sigma^2, \beta) p(\sigma^2) \\
 &\propto \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N \exp \left(-\frac{1}{2} \sum_{i=1}^N \frac{(y_i - x_i\beta)^2}{\sigma^2} \right) \times \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} \exp \left(-\frac{b}{\sigma^2} \right) \\
 &\propto (\sigma^2)^{(-a-1-\frac{N}{2})} \exp \left(-\frac{b}{\sigma^2} - \frac{1}{2} \sum_{i=1}^N \frac{(y_i - x_i\beta)^2}{\sigma^2} \right)
 \end{aligned}
 \tag{8}$$

IG(0, b)

(keep terms that are related to σ^2)

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

How

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

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

- Where

$$\begin{aligned}\bar{a} &= a + \frac{N}{2} \\ \bar{b} &= b + \frac{1}{2} \sum_{i=1}^N (y_i - x_i \beta)^2\end{aligned}\tag{9}$$

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 $s = 1$

1 $(\sigma^2)^s \sim p(\sigma^2 | Y, \beta^{(s-1)})$

2 $\beta^s \sim p(\beta | Y, \sigma^{(s-1)})$

- 3 Go to step 1 with $i = i + 1$ if $i < \underline{S}$. Otherwise, exit loop

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

Gibbs Sampling

Example: Linear regression

- Under regular conditions,

LLN

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

- 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

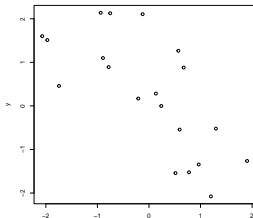
- ▶ Like before, I generate data y_i and x_i with

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

with

- ▶ $N = 100$
- ▶ $\beta_{\text{true}} = -1$ 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 $\beta^{(0)} = 5$

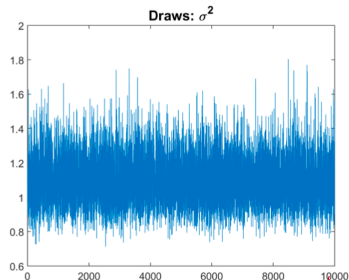
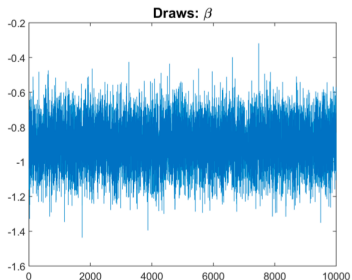
$N(0, 1)$
 $IG(10, 20)$



Gibbs Sampling

Example: Linear regression

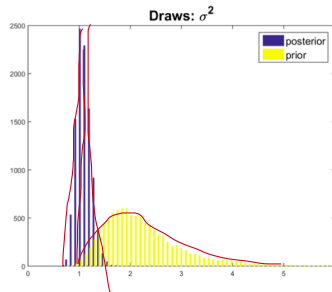
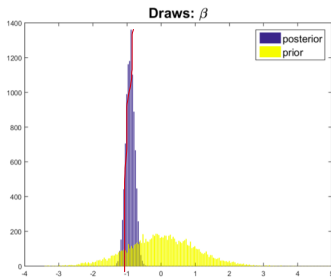
► Gibbs sampling: Posterior draws



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:

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

$$\sigma^2 = 1$$

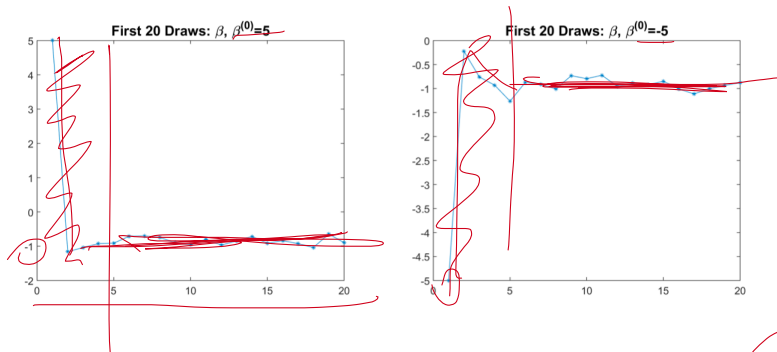
Lecture 9

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ (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)

Gibbs Sampling

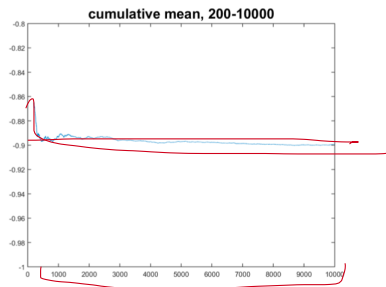
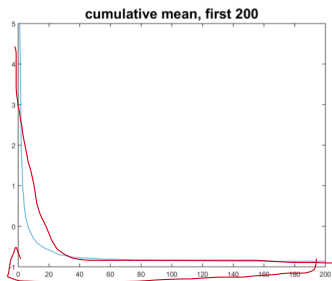
Checking convergence

- ▶ 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, $\beta^{(i)}, (\sigma^2)^i$ is a draw from $\pi(\beta, \sigma^2 | Y)$
- ▶ How can we check the convergence? There are many ways.

Gibbs Sampling

Checking convergence: Running Mean Plots

► Cumulative mean over draws

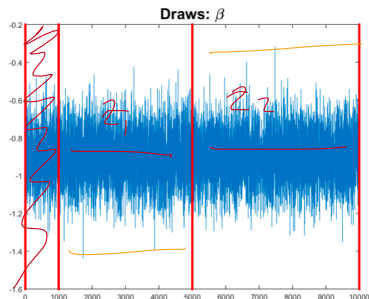


- This sequence behaves well in the sense that the Monte Carlo average converges to some number (posterior moment of interest) as the number of draws increases.

Gibbs Sampling

Checking convergence: Geweke's diagnostic check

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



- ▶ 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).

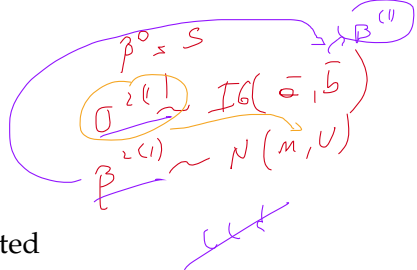
$$y = X\beta + u$$
$$u \sim N(0, \sigma^2)$$
$$\beta \sim N(0, \tau^2)$$
$$\sigma^2 \sim IG(a, b)$$

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

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

Gibbs Sampling

Gibbs sampling versus Direct sampling



- ▶ 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.

$$\begin{aligned} \tau^2 &\sim N(m, V) \\ \sigma^2 &\sim \text{IG}(\bar{a}, \bar{b}) \end{aligned}$$

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

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

close possible

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

Gibbs sample

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

- ▶ $V_\pi < V_G$
- ▶ 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 $\theta = (\theta_1, \theta_2, \dots, \theta_m)$
 - ▶ For each s it is possible to generate draws of θ_s from the conditional distribution, $p(\theta_s | \theta_{-s}, Y)$ where θ_{-s} denotes the vector θ without the partition θ_s
- ▶ Gibbs sampler: For $s = 1, \dots, S$: *2, n solo vic*
 - ▶ Draw $\theta_1^{(s+1)}$ from the density $p(\theta_1 | \theta_2^{(s)}, \theta_3^{(s)}, \dots, \theta_m^{(s)}, Y)$
 - ▶ Draw $\theta_2^{(s+1)}$ from the density $p(\theta_2 | \theta_1^{(s+1)}, \theta_3^{(s)}, \dots, \theta_m^{(s)}, Y)$
 - ▶ \vdots
 - ▶ Draw $\theta_m^{(s+1)}$ from the density $p(\theta_m | \theta_1^{(s+1)}, \theta_2^{(s+1)}, \theta_3^{(s+1)}, \dots, \theta_{m-1}^{(s+1)}, Y)$ *S*

Review & Next Steps

- Next Class: Empirical Bayes

- Next Week: PS 2

Data capacitor

$$\begin{aligned} y_i &\sim \mathcal{N}(\mu, \sigma^2) \\ \beta &\sim \mathcal{N}(\mu, \tau^2) \\ \sigma^2 &\sim \chi^2 \end{aligned}$$

$$IG(0, b)$$

$(0, 20)$

STAN

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.