Gibbs Sampling and Data Augmentation

Frank Schorfheide

University of Pennsylvania

Gerzensee Ph.D. Course on Bayesian Macroeconometrics

May 28, 2019

Gibbs Sampler

- Suppose the parameter vector θ can be partitioned into $\theta = [\theta_1', \ldots, \theta_m']'$.
- For each j it is possible to generate draws of θ_j from the conditional distribution $p(\theta_j|\theta_{-j}, Y)$, where θ_{-j} denotes the vector θ without the partition θ_j.
- For $j = 1, \ldots, N$:
  1. Draw θ_1^{i+1} from the density $p(\theta_1|\theta_2^i, \ldots, \theta_m^i, Y)$.
  2. Draw θ_2^{i+1} from the density $p(\theta_2|\theta_1^{i+1}, \theta_3^i, \ldots, \theta_m^i, Y)$.
  3. $\cdots$
  4. Draw θ_m^{i+1} from the density $p(\theta_m|\theta_1^{i+1}, \ldots, \theta_{m-1}^{i+1}, Y)$. $\square$

Gibbs Sampler

- Gibbs samplers belong to the class of Markov chain Monte Carlo (MCMC) algorithms.
- For large N we obtain dependent draws from the posterior distribution of θ.
- To reduce the influence of the initialization of the sampler, it is common practice to discard the initial draws.
- Approximate the posterior expectations of $h(\theta)$ by Monte Carlo averages:

$$\widehat{\mathbb{E}[\theta]} = \frac{1}{N - N_0} \sum_{i=N_0+1}^{N} h(\theta^i) \xrightarrow{a.s.} \mathbb{E}[h(\theta)|Y]$$

provided $\mathbb{E}\Big[|\theta(\theta)|\Big|Y\Big] < \infty$.

Back to the Basic State-Space Model

- Consider

$$\begin{aligned} y_t &= \Psi s_t + u_t \quad \text{measurement equation} \\ s_t &= \Phi s_{t-1} + \epsilon_t \quad \text{state transition equation} \end{aligned}$$

  where $\epsilon_t \sim iidN(0, \Sigma)$ and $u_t \sim iidN(0, H)$.

- y_t's are observed.

- s_t's are unobserved.

- Model generates joint density for the observations and latent states:

$$\begin{aligned} p(Y_{1:T}, S_{1:T}|\theta) &= \prod_{t=1}^{T} p(y_t, s_t|Y_{1:t-1}, S_{1:t-1}, \theta) \\ &= \prod_{t=1}^{T} p(y_t|s_t, \theta) p(s_t|s_{t-1}, \theta). \end{aligned}$$

Application: Time-varying Coefficients

- Consider the following model of inflation:

$$\pi_t = \pi_t^* + \tilde{\pi}_t$$

where π_t^* is a time-varying inflation target:

$$\tilde{\pi}_t = \rho\tilde{\pi}_{t-1} + \sigma_\epsilon \epsilon_t, \quad \pi_t^* = \pi_{t-1}^* + \sigma_\eta \eta_t.$$

- This looks like a state-space model:

$$\begin{aligned} y_t &= \begin{bmatrix} 1 & 1 \end{bmatrix} s_t \\ s_t &= \begin{bmatrix} \pi_t^* \\ \tilde{\pi}_t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \rho \end{bmatrix} s_{t-1} + \begin{bmatrix} \sigma_\eta & 0 \\ 0 & \sigma_\epsilon \end{bmatrix} \begin{bmatrix} \eta_t \\ \epsilon_t \end{bmatrix}. \end{aligned}$$

Application: Factor Models

- Extract "true" GDP growth from income and expenditure-side GDP measures.
- Measurement equation:

$$\begin{bmatrix} GDP_{Et} \\ GDP_{It} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} GDP_t + \begin{bmatrix} \epsilon_{Et} \\ \epsilon_{It} \end{bmatrix}$$

- State-transition equation:

$$GDP_t = \mu(1-\rho) + \rho GDP_{t-1} + \epsilon_{Gt}.$$

- We can also allow for correlation between measurement errors and state-transition innovations:

$$(\epsilon_{Gt}, \epsilon_{Et}, \epsilon_{It})' \sim iid\, N(\underline{0}, \Sigma), \quad \text{where} \quad \Sigma = \begin{bmatrix} \sigma^2_{GG} & 0 & 0 \\ 0 & \sigma^2_{EE} & \sigma^2_{EI} \\ 0 & \sigma^2_{IE} & \sigma^2_{II} \end{bmatrix}.$$

Bayesian Inference

- Suppose that all the non-redundant parameters of the state space model are collected in the vector θ.
- Bayes Theorem:

$$p(\theta|Y) \propto p(Y|\theta)p(\theta)$$

- We have learned how to numerically evaluate $p(Y|\theta)$ using the Kalman filter.
- But, how should we draw from the posterior?

Data Augmentation

- In the Bayesian framework, there is no conceptual difference between:
  - unknown model parameters θ,
  - latent states $S_{1:T}$.
- Implement a posterior sampler on the enlarged probability space for $(S_{1:T}, \theta)$.
- Bayes Theorem again:

$$p(\theta, S_{1:T} | Y_{1:T}) \propto \left(\prod_{t=1}^{T} p(y_t | s_t, \theta) p(s_t | s_{t-1}, \theta) \right) p(\theta)$$

- Construct a Gibbs sampler that iterates over parameters and states $S_{1:T}$:

$$\begin{aligned} p(S_{1:T} | Y_{1:T}, \theta) &\propto p(S_{1:T} | \theta) p(Y_{1:T} | S_{1:T}, \theta) \\ p(\theta | Y_{1:T}, S_{1:T}) &\propto p(\theta) p(S_{1:T} | \theta) p(Y_{1:T} | S_{1:T}, \theta) \end{aligned}$$

Bayesian Estimation of State-Space Model: Carter and Kohn (1994)

- Gibbs-sampling algorithm iterates over the conditional posteriors of θ and $S_{1:T}$.
- Recall the linear Gaussian state space representation

$$
\begin{aligned}
y_t &= A + Bs_t + u_t, \quad u_t \sim N(0, H) \\
s_t &= \Phi s_{t-1} + e_t, \quad e_t \sim N(0, Q)
\end{aligned}
$$

  with $\theta = (A, B, H, \Phi, Q)$
- For $i = 1, ..., n_{sim}$
  (a) Draw $\theta^{(i)}$ from $p\left(\theta \mid Y_{1:T}, S_{1:T}^{(i-1)}\right)$
    - Conditional on $S_{1:T}^{(i-1)}$, drawing θ is a standard linear regression
    - (Measurement) $y_t = A + Bs_t + u_t$
    - (Transition) $s_t = \Phi s_{t-1} + e_t$

  (b) Draw $S_{1:T}^{(i)}$ from $p\left(S_{1:T} \mid Y_{1:T}, \theta^{(i)}\right)$
    - Kalman / simulation smoother

Gibbs Sampler – Some Intuition

- Suppose we iterate over

$$p(\theta|\phi), \quad p(\phi|\theta).$$

- Define marginals

$$p(\theta) = \int_\Phi p(\theta|\phi)p(\phi)d\phi, \quad p(\phi) = \int_\Theta p(\phi|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}.$$

- Combine:

$$p(\theta) \;=\; \int_\Phi p(\theta|\phi)\left[\int_\Theta p(\phi|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}\right]d\phi = \int_\Theta\left[\int_\Phi p(\theta|\phi)p(\phi|\tilde{\theta})d\phi\right]p(\tilde{\theta})d\tilde{\theta}$$

- Define Markov transition kernel:

$$K(\theta|\tilde{\theta}) = \int_\Phi p(\theta|\phi)p(\phi|\tilde{\theta})d\phi$$

Gibbs Sampler – Some Intuition

- Recall Markov transition kernel:

$$K(\theta|\tilde{\theta}) = \int_{\Phi} p(\theta|\phi)p(\phi|\tilde{\theta})d\phi$$

- Note that $p(\theta)$ is a fixed point of the mapping $M[\cdot]$:

$$p(\theta) = \int K(\theta|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta} = M[p(\tilde{\theta})]$$

- Questions (see Tanner and Wong (1987) for answers):
  - Is the fixed point unique? Yes
  - Is $M[\cdot]$ a contraction mapping? Yes

Some Regularity Conditions

- $K(\theta|\tilde{\theta})$ is uniformly bounded and equicontinuous in θ.
- For any $\theta_0 \in \Theta$ there is a neighborhood $U(\theta_0)$ such that $K(\theta|\tilde{\theta}) > 0$ for all $\theta, \tilde{\theta} \in U(\theta_0)$.

Some Notation

- For a function $f(\theta)$ let $\|f\| = \int |f(\theta)| d\theta$.
- Recall the map $M[f] = \int K(\theta|\tilde{\theta}) f(\tilde{\theta}) d\tilde{\theta}$. Note that $M[\cdot]$ can be applied to a large class of functions $f(\cdot)$ (not just densities).

Useful Results

Lemma 1

Every fixed point of $M[\cdot]$ must be continuous.

- Let $p_*(\theta)$ be a fixed point of $M[\cdot]$.
- Consider

$$\begin{aligned}
&\lim_{\theta_1 \longrightarrow \theta_0} |p_*(\theta_1) - p_*(\theta_0)| \\
&= \lim_{\theta_1 \longrightarrow \theta_0} \left| \int K(\theta_1|\tilde{\theta}) p_*(\tilde{\theta}) d\tilde{\theta} - \int K(\theta_0|\tilde{\theta}) p_*(\tilde{\theta}) d\tilde{\theta} \right| \\
&\leq \lim_{\theta_1 \longrightarrow \theta_0} \int \left| K(\theta_1|\tilde{\theta}) - K(\theta_0|\tilde{\theta}) \right| p_*(\tilde{\theta}) d\tilde{\theta} \\
&= \int \left[\lim_{\theta_1 \longrightarrow \theta_0} \left| K(\theta_1|\tilde{\theta}) - K(\theta_0|\tilde{\theta}) \right| \right] p_*(\tilde{\theta}) d\tilde{\theta} \\
&= 0
\end{aligned}$$

- The second-to-last equality follows from the assumptions.

Useful Results

Lemma 2

$\|M[|f|]\| = \|f\|$

- Note that

$$
\begin{aligned}
\|M[|f|]\| &= \int_{\Theta}\left[\int_{\tilde{\Theta}} K(\theta|\tilde{\theta})|f(\tilde{\theta})|d\tilde{\theta}\right] d\theta \\
&= \int_{\tilde{\Theta}}\left[\int_{\Theta} K(\theta|\tilde{\theta})d\theta\right] |f(\tilde{\theta})|d\tilde{\theta} \\
&= \int_{\tilde{\Theta}} |f(\tilde{\theta})|d\tilde{\theta} \\
&= \|f\|
\end{aligned}
$$

Useful Results

Lemma 3

$\|M[f]\| \leq \|f\|$

- Note that

$$\begin{aligned} \|M[f]\| &= \int |M[f]| d\theta \\ &\leq \int M[|f|] d\theta \\ &= \|M[|f|]\| \\ &= \|f\| \end{aligned}$$

Useful Results

Lemma 4

Let $f^+ = f\{f \geq 0\}$ and $f^- = (-f)\{f < 0\}$. If f is such that neither f^+ nor f^- are identical to zero, then $\|M[f]\| < \|f\|$.

- Recall that $M[f] = \int K(\theta|\tilde{\theta})f(\tilde{\theta})d\tilde{\theta}$.
- Now consider

$$M[|f|] = M[f^+ + f^-] = M[f^+] + M[f^-], \quad |M[f]| = |M[f^+] - M[f^-]|.$$

- Note: $supp(f^+) \subset supp(M[f^+])$ and $supp(f^-) \subset supp(M[f^-])$.
- Deduce $supp(M[f^+])$ and $supp(M[f^-])$ overlap.
- In turn, $|M[f^+] - M[f^-]| < M[f^+ + f^-] = M[f^+] + M[f^-]$.
- Thus, $\|M[f]\| < \|M[|f|]\| = \|f\|$ (Lemma 2).

Uniqueness of Fixed Point

Uniqueness

p_* is the only density that satisfies $p_* = M[p_*]$.

- Suppose (to the contrary) $p_{**} = M[p_{**}]$ and define $f = p_* - p_{**}$.
- Then $M[f] = M[p_* - p_{**}] = p_* - p_{**} = f$ and f is a fixed point.
- f must be continuous (Lemma 1).
- Since $\int f(\theta)d\theta = 0$ and $f(\theta) \neq 0$ neither f^+ nor f^- can be zero.
- Thus, $\|M[f]\| < \|f\|$ (Lemma 4), which contradicts that f is a fixed point.

We want that $\|p_{(s+1)} - p_*\| < \|p_{(s)} - p_*\|$, where $p_{(s+1)} = M[p_{(s)}]$.

- It is straightforward to show the weaker result: $\|p_{(s+1)} - p_*\| \leq \|p_{(s)} - p_*\|$.
- Let $f = p_{(s)} - p_*$ such that $M[f] = p_{(s+1)} - p_*$.
- Desired result follows from Lemma 3 which states that $\|M[f]\| \leq \|f\|$.
- One can use arguments similar to those on the previous slide to turn the weak inequality into a strict inequality.

Convergence Rates

- Suppose that the starting value $p_{(0)}(\theta)$ satisfies $\sup_\theta \; p_{(0)}(\theta)/p_*(\theta) < \infty$.
- Then there exists a constant $\alpha \in (0,1)$ such that

$$\|p_{(s)} - p_*\| \le \alpha^s \|p_{(0)} - p_*\|$$

- See Tanner and Wong (1987).

Summary

- Let $p(\theta)$ be a normalized probability density. Define the mapping

$$M[p(\theta)] = \int K(\theta|\tilde{\theta}, Y)p(\tilde{\theta})d\tilde{\theta}$$

  $M[\cdot]$ maps a density $p(\theta)$ into a density $p'(\theta)$.
- We are interested in applying the mapping iteratively: Let $p^i(\theta) = M[p^{i-1}(\theta)]$.
- The mapping is constructed such that the fixed point corresponds to the posterior of interest.
- Under suitable regularity conditions
  1. The fixed point $p_*(\theta)$ of the mapping $M[\cdot]$ is unique.
  2. The mapping $M[\cdot]$ is a contraction mapping and the sequence of densities $\{p^i(\theta)\}_{i=0}^{\infty}$ converges to the fixed point $p_*(\theta)$

$$\int |p^i(\theta) - p_*(\theta)|d\theta \longrightarrow 0$$

     as $i \longrightarrow \infty$. $\square$

This Leads to .... Gibbs Sampler

- For $i = 1, \ldots, N$:
  1. Draw ϕ^{i+1} from the density $p(\phi|\theta^{i})$.
  2. Draw θ^{i+1} from the density $p(\theta|\phi^{i+1})$.

- It turns out that for $s > \bar{S}$ the marginal distribution of the draws (θ^{i}, ϕ^{i}) is approximately equal to the target distribution $p(\theta, \phi)$.

- However, the sequence of draws is serially correlated!

- Gibbs sampler creates a Markov chain. It belongs to the class of Markov chain Monte Carlo (MCMC) procedures.

More Generally

- Suppose the parameter vector θ can be partitioned into $\theta = [\theta_1', \dots, \theta_m']'$.
- For each j it is possible to generate draws of θ_j from the conditional distribution $p(\theta_j | \theta_{-j}, Y)$, where θ_{-j} denotes the vector θ without the partition θ_j.
- For $j = 1, \dots, N$:
  1. Draw θ_1^{i+1} from the density $p(\theta_1 | \theta_2^i, \dots, \theta_m^i, Y)$.
  2. Draw θ_2^{i+1} from the density $p(\theta_2 | \theta_1^{i+1}, \theta_3^i, \dots, \theta_m^i, Y)$.
  3. $\cdots$
  4. Draw θ_m^{i+1} from the density $p(\theta_m | \theta_1^{i+1}, \dots, \theta_{m-1}^{i+1}, Y)$. $\square$

Ergodicity

- A stationary process $\{\theta^i\}$ is said to be ergodic, if for any two bounded and measurable functions $f(\cdot)$ and $g(\cdot)$:

$$\lim_{n \longrightarrow \infty} \left| \mathbb{E}\big[f(\theta^i, \ldots, \theta^{i+k}) g(\theta^{i+n}, \ldots, \theta^{i+n+l})\big] \right| - \left| \mathbb{E}\big[f(\theta^i, \ldots, \theta^{i+k})\big] \right| \cdot \left| \mathbb{E}\big[g(\theta^{i+n}, \ldots, \theta^{i+n+l})\big] \right| = 0.$$

- If $\{\theta^i\}$ is strictly stationary and ergodic with $\mathbb{E}[|h(\theta)|] < \infty$, then

$$\frac{1}{N} \sum_{i=1}^{N} h(\theta^i) \xrightarrow{a.s.} \mathbb{E}[h(\theta)].$$

A Sufficient Condition for Ergodicity

Suppose that for every $\theta \in \Theta$ and every $A \subseteq \Theta$

$$\int_A p(\theta|Y)d\theta > 0 \quad \text{implies} \quad \int_A K(\tilde{\theta}|\theta)d\tilde{\theta} > 0$$

then the transition kernel of the Gibbs sampler is ergodic. (Geweke, 2005, Corollary 4.5.1)

Another Sufficient Condition for Ergodicity

Suppose that the following three conditions are satisfied:

- For all θ with $p(\theta|Y) > 0$ there exists an open neighborhood $N_\delta(\theta)$ such that for all $\tilde{\theta} \in N_\delta(\theta)$ $p(\tilde{\theta}|Y) > 0$.
- For every point $\tilde{\theta} \in \Theta$ and each block b of the Gibbs sampler, there exists an open neighborhood $N_\delta(\tilde{\theta}_{-b})$ of $\tilde{\theta}_{-b}$ and a bounded function $c(\tilde{\theta}_{-b})$ such that for all $\theta_{-b} \in N_\delta(\tilde{\theta}_{-b})$

$$\int_{\Theta(b)} p(\tilde{\theta}_{<b}, \theta_b, \tilde{\theta}_{>b}) d\theta_b \leq c(\tilde{\theta}_{-b})$$

- Θ is connected.

Then the transition kernel of the Gibbs sampler is ergodic. (Geweke, 2005, Theorem 4.5.4)

Gibbs Sampler

- For large N we obtain dependent draws from the posterior distribution of θ. It is common practice to discard the initial draws.

- Approximate the mean and covariance matrix of θ by Monte Carlo averages:

$$\widehat{\mathbb{E}[\theta]} = \frac{1}{N - N_0} \sum_{i=N_0+1}^{N} h(\theta^i) \xrightarrow{a.s.} \mathbb{E}[h(\theta)|Y]$$

provided $\mathbb{E}\big[|\theta(\theta)|\big|Y\big] < \infty$.

- Stronger regularity conditions are required to obtain a Central Limit Theorem (CLT)

$$\sqrt{N - N_0}\big(\widehat{\mathbb{E}}[\theta|Y] - \mathbb{E}[\theta|Y]\big) \Longrightarrow N(0, V)$$

- A CLT facilitates the computation of numerical standard errors for Monte Carlo approximations.

Illustration 1

- Suppose that

$$\left[\begin{array}{c}\theta_1\\ \theta_2\end{array}\right] \sim N\left(\left[\begin{array}{c}\mu_1\\ \mu_2\end{array}\right], \left[\begin{array}{cc}\Sigma_{11}=1 & \Sigma_{12}=\rho\\ \Sigma_{21}=\rho & \Sigma_{22}=1\end{array}\right]\right)$$

- Conditional distribution 1:

$$\theta_1|\theta_2 \sim N\big(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\theta_2-\mu_2),\, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\big)$$

- Conditional distribution 2:

$$\theta_2|\theta_1 \sim N\big(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\theta_1-\mu_1),\, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\big)$$

- Vary ρ and observe performance.

Illustration 1

- If parameters are highly correlated across blocks the draws will also be highly correlated and the sampler moves slowly through the parameter space.
- What's bad about large serial correlation?

$$\sqrt{N}(\bar{h}_N - \mathbb{E}[\bar{h}_N])$$
$$\Longrightarrow \quad N\left(0, \frac{1}{N}\sum_{i=1}^{n}\mathbb{V}[h(\theta^i)] + \frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i} COV\left[h(\theta^i), h(\theta^j)\right]\right).$$

Illustration 2 - Dynamic Factor Model

- Observables:

$$y_{it} = \alpha_i + \lambda_i f_t + \xi_{it}$$

- Common factor:

$$f_t = \phi_0 f_{t-1} + u_{0t} \quad u_{0t} \sim N(0, \sigma^2)$$

- Idiosyncratic processes

$$\xi_{it} = \phi_i \xi_{it-1} + u_{it} \quad u_{it} \sim N(0, \sigma_i^2)$$

- Grouping of parameters: $\theta_0 = [\phi_0, \sigma^2]$, $\theta_{1i} = [\phi_1, \sigma_i^2, \lambda_i, \alpha_i]$.

Illustration 2 - Dynamic Factor Model

- Conditional on $(\alpha_i, \lambda_i, f_{1:T})$ we can compute

$$\xi_{it} = y_{it} - \alpha_i - \lambda_i f_t$$

and estimate a Bayesian regression model for

$$\xi_{it} = \phi_i \xi_{it-1} + u_{it}.$$

- Conditional on $(\phi_i, f_{1:T})$ we can estimate the quasi-differenced regression:

$$y_{it} - \phi_i y_{it-1} = \alpha_i \cdot (1 - \phi_i) + \lambda_i \cdot (f_t - \phi_i f_{t-1}) + u_{it}.$$

- Conditional on $f_{1:T}$ we can estimate the regression:

$$f_t = \phi_0 f_{t-1} + u_t.$$

Illustration 2 - Dynamic Factor Model

- State-space representation for filter/smoother...
- There are N measurement equations:

$$y_{it} - \phi_i y_{it-1} = \alpha_i \cdot (1 - \phi_i) + \lambda_i \cdot f_t - (\lambda_i \phi_i) \cdot f_{t-1} + u_{it}.$$

- State vector $s_t = [f_t, f_{t-1}]'$.
- State transition (companion form VAR):

$$\begin{bmatrix} f_t \\ f_{t-1} \end{bmatrix} = \begin{bmatrix} \phi_0 & 0 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} f_{t-1} \\ f_{t-2} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u_t.$$