14.384 Time Series Analysis, Fall 2007
Professor Anna Mikusheva
Paul Schrimpf, scribe
December 11, 2007

Lecture 26

MCMC: Gibbs Sampling

Last time, we introduced MCMC as a way of computing posterior moments and probabilities. The idea was to draw a sample from the posterior distribution and use moments from this sample. We drew these samples by constructing a Markov Chain with the posterior distribution as its invariant measure. In particular, we found a transition kernel, $P(x, dy)$, such that $\pi(y) = \int P(x, dy)\pi(x)dx$. The Gibbs sampler is a special case of MCMC.

Gibbs Sampling

Suppose we can write our random variable of interest as components, $x = (x_1, x_2, ..., x_d)$, such that we can simulate the distribution of each component conditional on the others, i.e. we can draw from $\pi(x_k|x_1, ..., x_{k-1}, x_{k+1}, ..., x_d)$ $\forall k$. We want to sample from the joint distribution, $\pi(x)$. Gibbs sampling constructs a Markov Chain, $x^{(1)} \to x^{(2)} \to ...$, with step $x^{(j)} \to x^{(j+1)}$ given by:
Simulate

- $x_1^{(j+1)}$ from $\pi(x_1^{(j+1)}|x_2^{(j)}, ..., x_d^{(j)})$
- $x_2^{(j+1)}$ from $\pi(x_2^{(j+1)}|x_1^{(j+1)}, x_3^{(j)}, ..., x_d^{(j)})$
- $x_3^{(j+1)}$ from $\pi(x_3^{(j+1)}|x_1^{(j+1)}, x_2^{(j+1)}, x_4^{(j)}, ..., x_d^{(j)})$
- ...

Claim 1. $\pi(x)$ is the invariant measure for this Markov Chain.

Proof. The transition kernel is:

$$\begin{aligned}P(x, dy) =& \pi(y_1|x_2, ..., x_d)\pi(y_2|y_1, x_3, ..., x_d)...\pi(y_d|y_1, ..., y_{d-1})dy_1...dy_d \\ =& \prod_k \pi(y_k|y_1, ..., y_{k-1}, x_{k+1}, ..., x_d)dy_k\end{aligned}$$

We want to show that $\int P(x, dy)\pi(x)dx = \pi(y)dy$. Consider:

$$\begin{aligned}\int P(x, y)\pi(x)dx =& \int \prod_k \pi(y_k|y_1, ..., y_{k-1}, x_{k+1}, ..., x_d)\pi(x_1, ..., x_d)dx_1...dx_d \\ =& \text{Bayes rule} = \\ =& \int \prod \frac{\pi(y_k|y_1, ..., y_{k-1})\pi(x_{k+1}, ..., x_d|y_1, ..., y_k)}{\pi(x_{k+1}, ..., x_d|y_1, ..., y_{k-1})}\pi(x_1|x_2, ..., x_d)\pi(x_2, ..., x_d)dx_1...dx_d \\ =& \prod \pi(y_k|y_1, ..., y_{k-1}) \int \prod \frac{\pi(x_{k+1}, ..., x_d|y_1, ..., y_k)}{\pi(x_{k+1}, ..., x_d|y_1, ..., y_{k-1})}\pi(x_1|x_2, ..., x_d)\pi(x_2, ..., x_d)dx_1...dx_d\end{aligned}$$

Then, since $\prod \pi(y_k|y_1,...,y_{k-1}) = \pi(y)$, we need to show that

$$\begin{aligned}
1 =& \int \prod \frac{\pi(x_{k+1},...,x_d|y_1,...,y_k)}{\pi(x_{k+1},...,x_d|y_1,...,y_{k-1})}\pi(x_1|x_2,...,x_d)\pi(x_2,...,x_d)dx_1...dx_d \\
=& \int \frac{\pi(x_2,...,x_d|y_1)}{\pi(x_2,...,x_d)}\frac{\pi(x_3,...,x_d|y_1,y_2)}{\pi(x_3,...,x_d|y_1)}\frac{\pi(x_4,...,x_d|y_1,...,y_3)}{\pi(x_4,...,x_d|y_1,y_2)}...\frac{\pi(x_d|y_1,...,y_d)}{\pi(x_d|y_1,...,y_{d-1})}\pi(x_1|x_2,...,x_d)\pi(x_2,...,x_d)dx_1...dx_d \\
=& \int \prod \frac{\pi(x_{k+1},...,x_d|y_1,...,y_k)}{\pi(x_{k+2},...,x_d|y_1,...,y_k)}\pi(x_1|x_2,...,x_d)dx_1...dx_d \\
=& \int \prod \pi(x_k|x_{k+1},...,x_d,y_1,...,y_{k-1})dx_1...dx_d
\end{aligned}$$

Then we start with the integral over dx_1: x_1 enters only $\pi(x_1|x_2,...,x_d)$. We have $\int \pi(x_1|x_2,...,x_d)dx_1 = 1$. Then we notice that after the first integration x_2 enters only $\pi(x_2|x_3,...,x_d,y_1)$ and $\int \pi(x_2|x_3,...,x_d,y_1)dx_2 = 1$, and so on. □

Questions about Gibbs sampling

- *How it relates to Metropolis-Hastings?* It's a special case. You can calculate that the acceptance rate is equal to 1.
- Why would one use Gibbs sampling? 1) It can be used in many cases when the functional form of the density is not analytical. 2) It allows for data augmentation (and inferences about hidden states).

Data Augmentation

Example: Tobit

Suppose $z_i \sim N(x_i'\beta, \sigma^2)$, $y_i = \max\{0, z_i\}$. For simplicity, let $\sigma^2 = 1$. We observe $\mathcal{Y}_T = \{y_1, ..., y_T\}$ and $\{x_t\}$ but not $\{z_t\}$. Let $\mathcal{C} = \{i : y_i = 0\}$ be the set of censored observations. The likelihood is:

$$\prod_{i\in\mathcal{C}}(1-\Phi(x_i'\beta))\prod_{i\notin\mathcal{C}}\phi(y_i - x_i'\beta)$$

Notice, that the likelihood is not analytical (since Φ is some integral that does not have a closed form). Let the prior for β be $\beta \sim N(\beta_0, B_0)$. What is the posterior of $\beta|\mathcal{Y}_T$? The problem of running a regular Metropolis-Hastings is in non-analytical form of likelihood (and as a result, it is difficult to calculate the rejection probability).

A solution is to use Gibbs sampling and **data augmentation**. The data augmentation idea is to increase the parameter space by adding hidden states $\tilde{Z} = \{z_i\}_{i\in\mathcal{C}}$. The idea is to simulate from the joint distribution of $\tilde{Z} = \{z_i\}_{i\in\mathcal{C}}$ and β given $\mathcal{Y}_T$.

For Gibbs sampling we have to be able to simulate from the following two conditional densities: 1) $\beta|\tilde{Z}, \mathcal{Y}_T$; and 2) $\tilde{Z}|\beta, \mathcal{Y}_T$. We consider β to be the first block x_1, and $\tilde{Z}$ as the second block x_2. Then by simulating $\beta^{(t+1)}|\tilde{Z}^{(t)}, \mathcal{Y}_T$ and $\tilde{Z}^{(t+1)}|\beta^{(t+1)}, \mathcal{Y}_T$, we will get (at the limit) a draw from the joint distribution $\beta, \tilde{Z}|\mathcal{Y}_T$. If then we discard the $\tilde{Z}$ part of the draw, we have a draw from $\beta|\mathcal{Y}_T$.

We can notice that if the state ($\tilde{Z}$) is observable then we have a regular OLS model: $z_i = x_i'\beta + \varepsilon_i, \ \ \varepsilon_i \sim N(0,1)$ (we observe now both x_i and z_i: $z_i = y_i$ if $i \notin \mathcal{C}$ and $z_i \in \tilde{Z}$ if $i \in \mathcal{C}$). As a result $\beta|\tilde{Z}, \mathcal{Y}_T$ is normal and easy to simulate. In particular the conditional distributions of β is (we derived it last time for a case with $\beta_0 = 0$ and $B_0 = \tau I_k$):

$$\begin{aligned}
\pi(\beta|\tilde{Z}, \mathcal{Y}_T) =& \pi(\beta|\{z_i\}) \sim N(\tilde{\beta}, \tilde{B}) \\
\tilde{\beta} =& (B_0 + X'X)^{-1}(B_0\beta_0 + X'Z) \\
\tilde{B} =& (B_0 + X'X)^{-1}
\end{aligned}$$

Note, that the values of $\tilde{\beta}$ and $\tilde{B}$ depend on the full vector of Z, not only on $\mathcal{Y}_T$.

We also know that the distribution of $\tilde{Z}|\beta, \mathcal{Y}_T$ is easy to simulate from. Remember that $\tilde{Z}$ include only censored(unobserved values of $z_i < 0$), as a result, $\tilde{Z}$ is independent on $\mathcal{Y}_T$. The conditional distribution of $\tilde{Z}$ is

$$\begin{aligned}\pi(\tilde{Z}|\beta, \mathcal{Y}_T) =& \prod_{i\in\mathcal{C}} f(z_i|\beta, y_i = 0)\\ =& \prod_{i\in\mathcal{C}} f(z_i|\beta, z_i \leq 0)\end{aligned}$$

This is a truncated normal distribution. One can draw it in two ways: 1) simulate error term $\varepsilon_i \sim N(0,1)$, calculate $z_i = x_i'\beta + \varepsilon_i$, if $z_i < 0$ keep it, otherwise, discard and repeat drawing of ε_i. Or 2) we can draw from it by taking:

$$z_i = x_i'\beta + \Phi^{-1}(\Phi(-x_i'\beta)u_i)$$

where $u_i \sim U[0,1]$. But here you should be able to calculate Φ and Φ^{-1} fast and accurately (that returns us back to why Metropolis- Hastings can suffer difficulties).

We can then use Gibbs sampling to simulate the joint distribution, $\tilde{Z}, \beta|\mathcal{Y}_T$. If we are only interested in β, we can just ignore the draws of $\tilde{Z}$.

Practical implementation, and convergence

Assume that we have a Markov chain X_t generater with a help of Metropolis-Hastings algorithm (Gibbs sampling is a special case of it). The starting point is X_0 and a transitional kernel $P(x, A)$. We are interested in convergence of the chain to its invariant distribution π, that is, we expect the distribution$P^{(t)}(A) = P\{X_t \in A|X_0\}$ to converge $P^{(t)} \Rightarrow \pi$.

The general Markov Chain theory gives us the following. A chain is called $\pi-$*irreducible* if for every x and for all A such that $\pi(A) > 0$ we have $P(X_t \in A|X_0 = x) > 0$ for some t. (That is, all sets are visited with positive probability). A chain is called aperiodic if there exists no partition $(D_0, D_1, ..., D_{p-1})$ for some $p > 1$ such that $P\{X_t \in D_{t \mod (p)}|X_0 \in D_0\} = 1$ for all t.

Theorem 2. *If $P(x, A)$ is $\pi-$irreducible transition kernel and has an invariant distribution π, then π is the unique invariant. If $P(x, A)$ is also aperiodic, then for $\pi-$almost all x and all sets A we have:*

$$|P\{X_t \in A|X_0 = x\} - \pi(A)| \to 0 \quad as \quad t \to \infty$$

If h is π integrable real-valued function, then

$$\frac{1}{T}\sum_{t=1}^{T} h(X_t) \to \int h(x)\pi(dx) \quad a.s., T \to \infty$$

For Metropolis-Hastings it also known that: 1) convergence is geometric for independence M-H ($q(x, y) = q(y)$) if q/π is bounded over the support; 2) convergence is geometric for random walk M-H if π has geometric tails.

To detect convergence in practice one may do the following:

- check whether some characteristics of the chain distribution (such as mean, variance, median, quartiles) are stabilize with time for a single long run;
- compare a ratios of a current estimate of a target density and a target density itself (known up to a constant) for a single long run of a chain;
- compare some characteristics of the chain distribution across several independent runs of a chain.

General practical recommendations:

- There is no unique way to implement MCMC. In majority of cases you have many choices, try several of them. It relates to choosing between different variants of M-H and Gibbs sampling as well as to picking up sampling density q. For better results q should be appropriately centered and have tails dominating π.
- I've seen a recommendation to tune your M-H algorithm to roughly 30% acceptance rate. I still don't know where the number comes from... The general wisdom is: if your acceptance rate is too low, then you are producing very dependent draws; if your acceptance rate is too high, it may be the sign that you are visiting only the high probability area and not moving too far from it.
- **Blocking** Parameters or state variables that are correlated should, if possible, be drawn in blocks. It can improve the convergence speed dramatically. If two parameters are highly correlated and drawn consequently in Gibbs sampling- they almost will not be updated in any single run. (see example below about diffusion)
- You should be very careful with choosing "non-informative" priors, since many of them tend to be improper and may produce improper posteriors (ooo-u-ps!)

Example: State-Space Model

Here we consider an unrestricted linear Gaussian state space model. Consider the model

$$(*) \quad \begin{aligned} y_t =& A + Bs_t + u_t \\ s_t =& \Phi s_{t-1} + \varepsilon_t \end{aligned}$$

where $E[u_t u_t'] = H$ and $E[\varepsilon_t \varepsilon_t'] = I$. We consider parameters A, B, H, Φ distributed with a prior distribution $\pi(A, B, H, \Phi)$. We have two ways to go- M-H and Gibbs sampling. Let's introduce $S_T = \{s_1, ..., s_T\}$, $Y_T = \{y_1, ..., y_T\}$.

Metropolis-Hastings:

$$\pi(A, B, H, \Phi | Y_T) \propto \pi(A, B, H, \Phi) L(Y_T | A, B, H, \Phi),$$

where $L(Y_T | A, B, H, \Phi)$ can be calculated using Kalman filter. It is straight- forward to run M-H.

Gibbs sampling Assume additionally that priors are the following: A, B are jointly normal, Φ is normal, H is known. We do data augmentation. For Gibbs sampling we have to be able to simulate from the following conditional distributions:

- $p(S_T | Y_T, A, B, \Phi, H)$: it is normal with parameters known from Kalman smoother.
- $p(A, B | Y_T, S_T, \Phi, H)$: notice that conditional on the states this problem is a standard multivariate regression
$$y_t = A + Bs_t + u_t, \quad \text{where} \ \ u_t \sim N(0, H),$$
so the posterior for A, B conditional on Y_T, S_T is normal (exact formulas are in all Bayesian books)
- $p(\Phi | Y_T, S_T, A, B, H)$ is based on the VAR (also normal)
$$s_t = \Phi s_{t-1} + \varepsilon_t, \quad \text{where} \ \ \varepsilon_t \sim N(0, I)$$

What pluses from using Gibbs sampling: we can get joint inference about latent states and parameters. Problem: special form of priors. The problem can be overcome by "Metropolis-Hastings inside Gibbs sampling".

Joining Gibbs and Metropolis-Hastings

We want to simulate from a joint density $\pi(x_1, x_2)$. Assume that for random variables $(\xi_1, \xi_2) \sim \pi(x_1, x_2)$ we have the following conditional distributions: $\pi(x_1|x_2)dx_1 = P(\xi_1 \in dx_1|\xi_2 = x_2)$ and $\pi(x_2|x_1)dx_2 = P(\xi_2 \in dx_2|\xi_1 = x_1)$

Theorem 3. *Let us have two transition kernels, $P_1(x_1, dy_1|x_2)$ with invariant $\pi(x_1|x_2)$ and $P_2(x_2, dy_2|x_1)$ with invariant distribution $\pi(x_2|x_1)$. Then $P(x, dy) = P_1(x_1, dy_1|x_2)P_2(x_2, dy_2|y_1)$ has invariant distribution $\pi(x_1, x_2)$.*

Proof.

$$\begin{aligned}
&\int\int P_1(x_1, dy_1|x_2)P_2(x_2, dy_2|y_1)\pi(x_1, x_2)dx_1dx_2 = \langle \pi(x_1, x_2) = \pi(x_1|x_2)\pi(x_2)\rangle \\
=&\int \left(\int P_1(x_1, dy_1|x_2)\pi(x_1|x_2)dx_1\right) P_2(x_2, dy_2|y_1)\pi(x_2)dx_2 \\
=&dy_1 \int \pi(y_1|x_2)P_2(x_2, dy_2|y_1)\pi(x_2)dx_2 = \langle \pi(y_1|x_2)\pi(x_2) = \pi(x_2|y_1)\pi(y_1)\rangle \\
=&\pi(y_1)dy_1 \int \pi(x_2|y_1)P_2(x_2, dy_2|y_1)dx_2 = \pi(y_1)dy_1\pi(y_2|y_1)dy_2 = \pi(y_1, y_2)dy
\end{aligned}$$

□

This procedure suggested above is called Metropolis-Hastings within Gibbs sampler. Assume that one can construct for each value x_2 a Markov chain (M-H) $x_1^{(t)}$ such that at the limit $x_1^{(t)}$ is distributed as $\pi(x_1|x_2)$ and for each value x_1 a Markov chain (M-H) $x_2^{(t)}$ such that at the limit $x_2^{(t)}$ is distributed as $\pi(x_2|x_1)$. We want construct a draw from the joint distribution $\pi(x_1, x_2)$. One can run a Gibbs sampling $x^{(1)} \to x^{(2)} \to ...$ with a single step of Metropolis-Hastings on every step. That is, we going from $x^{(t)} = (x_1^{(t)}, x_2^{(t)})$ to $x^{(t+1)} = (x_1^{(t+1)}, x_2^{(t+1)})$ in the following way:

- do one draw of Metropolis-Hastings by simulating $x_1^{(t+1)}$ using transition kernel $P(x_1^{(t)}, \cdot|x_2^{(t)})$
- do one draw of Metropolis-Hastings by simulating $x_2^{(t+1)}$ using $P(x_2^{(t)}, \cdot|x_1^{(t+1)})$

In example of state-space model (*) assume that we relax assumption of normality of A, B, instead assume that the prior for A, B is $\mu(A, B)$. Could we still use Gibbs sampling? Yes. This is how we will do one step in going from $(S_T^{(t)}, A^{(t)}, B^{(t)}, \Phi^{(t)})$ to $(S_T^{(t+1)}, A^{(t+1)}, B^{(t+1)}, \Phi^{(t+1)})$

- Simulate $S_T^{(t+1)}$ from $p(S_T|Y_T, A^{(t)}, B^{(t)}, \Phi^{(t)}, H)$: it is normal with parameters known from Kalman smoother (no change here).
- Simulate $A^{(t+1)}, B^{(t+1)}$ from $p(A, B|Y_T, S_T^{(t+1)}, \Phi^{(t)}, H)$: problem here - if prior for A, B is not normal, then the distribution may be intractable:

  $$p(A, B|Y_T, S_T, \Phi, H) \propto \mu(A, B)\prod_{t=1}^{T}(\det(H))^{-1/2}\exp\left\{-\frac{1}{2}(y_t - A - Bs_t)'H^{-1}(y_t - A - Bs_t)\right\} = f(A, B, Y_T, S_T)$$

  However, it can be simulated using M-H. We would do just one step of it.

  - Simulate a draw $(\widetilde{A}, \widetilde{B})$ from some density q (it may depend on any variable known)
  - calculate the probability of moving

    $$\alpha(A^{(t)}, B^{(t)}, \widetilde{A}, \widetilde{B}) = \min\left\{1, \frac{f(\widetilde{A}, \widetilde{B}, Y_T, S_T^{(t+1)})}{f(A^{(t)}, B^{(t)}, Y_T, S_T^{(t+1)})}\frac{q(A^{(t)}, B^{(t)})}{q(\widetilde{A}, \widetilde{B})}\right\}$$

    Note that the probability of moving depends on the current draw of S_T

  - Accept $(A^{(t+1)}, B^{(t+1)}) = (\widetilde{A}, \widetilde{B})$ with probability $\alpha(A^{(t)}, B^{(t)}, \widetilde{A}, \widetilde{B})$, other wise, use $(A^{(t+1)}, B^{(t+1)}) = (A^{(t)}, B^{(t)})$.

- Simulate $\Phi^{(t+1)}$ from $p(\Phi|Y_T, S_T^{(t+1)}, A^{(t+1)}, B^{(t+1)}, H)$ is based on the VAR (normal as before)

$$s_t = \Phi s_{t-1} + \varepsilon_t, \quad \text{where } \ \varepsilon_t \sim N(0, I)$$

Addendum to State-Space Model: how to use Kalman smoother

I am returning to a question on how to use Kalman smoother in state-space model to update state variables (to simulate from $p(S_T|Y_T, A, B, \Phi, H)$). Again we have a state-space model:

$$\begin{aligned} (*) \quad y_t =& A + Bs_t + u_t \\ s_t =& \Phi s_{t-1} + \varepsilon_t \end{aligned}$$

We want to derive the posterior for $S_T = \{s_t, t = 1, ..., T\}$, $p(S_T|Y_T, A, B, \Phi, H))$. We could use Gibbs sampling and sample each s_t separately. However, if Φ is near 1, then the s_t will be very correlated and Gibbs sampling will result in a highly auto-correlated, slow converging Markov Chain. Therefore, it is best to sample S_T as a block. Notice that:

$$\begin{aligned} p(S_T|Y_T, A, B, \Phi, H) =& \prod_t p(s_t|Y_T, A, B, \Phi, H, s_{t+1}, ..., s_T) \\ =& \prod_t p(s_t|Y_T, A, B, \Phi, H, s_{t+1}) \\ =& \prod_t p(s_t|Y_t, A, B, \Phi, H, s_{t+1}) \\ =& \prod_t \frac{p(s_t|Y_t, A, B, \Phi, H)p(s_{t+1}|Y_t, A, B, \Phi, H, s_t)}{p(s_{t+1}|Y_t, A, B, \Phi, H)} \end{aligned}$$

That is, we will sample S_T starting with time T and going to time 1. After s_{t+1} is drawn, the draw of s_t has conditional pdf proportional to

$$p(s_t|Y_t, A, B, \Phi, H)p(s_{t+1}|Y_t, A, B, \Phi, H, s_t)$$

Notice that the first term could be written from Kalman filter $\sim N(s_{t|t}, P_{t|t})$, and the second term, $p(s_{t+1}|Y_t, A, B, \Phi, H, s_t)$, is is our state equation. As a result,

$$\begin{aligned} p(s_t|Y_T, A, B, \Phi, H, s_{t+1}) \propto& \exp\left(\frac{-1}{2P_{t|t}}(s_t - s_{t|t})^2\right) \exp\left(\frac{-1}{2}(s_{t+1} - \Phi s_t)^2\right) \\ \propto& \exp\left(-\frac{1}{2\tilde{P}_t}(s_t - \tilde{s}_t)^2\right) \end{aligned}$$

where $\tilde{s}_t$ and $\tilde{P}_t$ can be calculated from $s_{t|t}$, $P_{t|t}$, s_{t+1}, and Φ. So to update a value of S_T in Gibbs sampler we proceed as follows:

- run Kalman filter to get $s_{t|t}$,$P_{t|t}$
- $s_T \sim N(s_{T|T}, P_{T|T})$
- For $t = T, T-1, ..., 1$:
  - Use $s_{t-1|t-1}$, $P_{t-1|t-1}$,s_t to calculate $\tilde{s}_{t-1}$, $\tilde{P}_{t-1}$
  - draw $s_{t-1} \sim N(\tilde{s}_{t-1}, \tilde{P}_{t-1})$

14.384 Time Series Analysis
Fall 2013