14.384 Time Series Analysis, Fall 2007
Professor Anna Mikusheva
Paul Schrimpf, scribe
Novemeber 29, 2007

Lecture 25

MCMC: Metropolis Hastings Algorithm

A good reference is Chib and Greenberg (*The American Statistician* 1995).
Recall that the key object in Bayesian econometrics is the posterior distribution:

$$p(\theta|\mathcal{Y}_T) = \frac{f(\mathcal{Y}_T|\theta)p(\theta)}{\int f(\mathcal{Y}_T|\tilde{\theta})d\tilde{\theta}}$$

It is often difficult to compute this distribution. In particular, the integral in the denominator is difficult. So far, we have gotten around this by using conjugate priors – classes of distributions for which we know the form of the posterior. Generally, it's easy to compute the numerator, $f(\mathcal{Y}_T|\theta)p(\theta)$, but it is hard to compute the normalizing constant, the integral in the denominator, $\int f(\mathcal{Y}_T|\tilde{\theta})d\tilde{\theta}$. One approach is to try to compute this integral in some clever way. Another, more common approach is Markov Chain Monte-Carlo (MCMC). The goal here is to generate a random sample $\theta_1, .., \theta_N$ from $p(\theta|\mathcal{Y}_T)$. We can then use moments from this sample to approximate moments of the posterior distribution. For example,

$$E(\theta|\mathcal{Y}_T) \approx \frac{1}{N}\sum \theta_n$$

There are a number of methods for generating random samples from an arbitrary distribution.

Acceptance-Rejection Method (AR)

We start with the simplest one. The goal is to simulate $\xi \sim \pi(x)$.

What we know: 1) a function, $f(x)$, such that $\pi(x) = \frac{f(x)}{k}$ The constant k is unknown (that is, f is a pdf up to an unknown normalization). 2)we can simulate draws from some candidate pdf $h(x)$; 3)there is a known constant c such that $f(x) \leq ch(x)$

We simulate draws from $\pi(x)$ as follows:

1. Draw $z \sim h(x)$, $u \sim U[0,1]$
2. If $u \leq \frac{f(z)}{ch(z)}$, accept the draw $\xi = z$. Otherwise discard the draw and repeat (1)

The intuition of the procedure is the following: Let $v = uch(z)$ and imagine the joint distribution of (v, z). It has support under the graph of $ch(z)$ with a uniform density (it is uniform on $\{(v,z) : z \in Support(h), 0 \leq v \leq ch(z)\}$). Then, it is fairly easy to see that if we accept $\xi = z$, the joint distribution of (v, ξ) is uniform over the support $\{(v,\xi) : \xi \in Supportt(\pi), f(\xi) \geq v \geq 0\}$. Then (for the same reason that $h(z)$ is the marginal density of (v, z)), the marginal density of ξ will be $\frac{f(\xi)}{k}$. More formally,

Proof. Let ρ be the probability of rejecting a single draw. Then,

$$\begin{aligned}
P(\xi \leq x) =& P\left(z_1 \leq x, u_1 \leq \frac{f(z_1)}{ch(z_1)}\right)(1 + \rho + \rho^2 + ...) \\
=& \frac{1}{1-\rho}P\left(z_1 \leq x, u_1 \leq \frac{f(z_1)}{ch(z_1)}\right) = \frac{1}{1-\rho}E_z\left[P\left(u \leq \frac{f(z)}{ch(z)}|z\right)\mathbf{1}_{\{z \leq x\}}\right] \\
=& \frac{1}{1-\rho}\int_{-\infty}^{x}\frac{f(z)}{ch(z)}h(z)dz = \int_{-\infty}^{x}\frac{f(z)}{c(1-\rho)}dz = \int_{-\infty}^{x}\pi(z)dz
\end{aligned}$$

The last line is due to the fact that there exists the unique constant that normalizes f to be a pdf. Since the left hand side is a cdf, then $\frac{1}{c(1-\rho)}$ is this constant. □

A major drawback of this method is that is may lead us to reject many draws before we finally accept one. This can make the procedure inefficient. If we choose c and $h(z)$ poorly, then $\frac{f(z)}{ch(z)}$ could be very small for many z. It will be especially difficult to choose a good c and $h()$ when we do not know much about $\pi(z)$.

Markov Chains

A Markov Chain is a stochastic process where the distribution of x_{t+1} only depends on x_t, $P(x_{t+1} \in A|x_t, x_{t-1}, ...) = P(x_{t+1} \in A|x_t)\ \forall A$.

Definition 1. A *transition kernel* is a function, $P(x, A)$, such that, for every x it is a probability measure in the second argument:

$$P(x, A) = P\left(x_{t+1} \in A|x_t = x\right)$$

It gives the probability of moving from x into the set A.

The transition kernel may have atoms, in particular, we would be considering cases with non-zero probability of (not moving) staying: $P(x, \{x\}) \neq 0$.

We want to study the behavior of a sequence of draws $x_1 \to x_2 \to ...$ where we move around according to a transition kernel. Suppose the distribution of x_t is $P^{(t)}$, then the distribution of $y = x_{t+1}$ is

$$P^{(t+1)}(y)dy = \int_{\Re} P^{(t)}(x)P(x, dy)dx.$$

Definition 2. A distribution π^* is called an*invariant measure* (with respect to transition kernel $P(x, A)$) if $\pi^*(y)dy = \int_{\Re} \pi^*(x)P(x, dy)dx$.

Under some regularity conditions, a transition kernel $P(x, A)$ has a unique invariant distribution π^*; and a marginal distribution $P^{(t)}$ of x_t - an element in Markov chain with the transitional kernel $P(x, A)$ converges to its invariant distribution π^* as $t \to \infty$. That is, if one would run a Markov chain long enough then the distribution of the draw is close to π^*. Generally, if the transition kernel is irreducible (it can reach any point from any other point) and aperiodic (not periodic, *i.e.* the greatest common denominator of $\{n : y$ can be reached from x in n steps$\}$ is 1), then it converges to an invariant distribution.

A classical Markov chain problem is to find π^* given $P(x, A)$. The MCMC has an inverse problem. Assume we want to simulate a draw from π^* (which we know up to a constant multiplier). We need to find a transition kernel $P(x, dy)$ such that π^* is its invariant measure. Let's suppose that π^* is continuous. We will consider the class of kernels

$$(*) \qquad P(x, dy) = p(x, y)dy + r(x)\Delta_x(dy),$$

here $\Delta_x(dy)$ is a unit mass measure concentrated at point x: $\Delta_x(A) = \mathbb{I}\{x \in A\}$. So, the transition kernel (*) says that we can stay at x with probability $r(x)$, otherwise y is distributed according to some pdf proportional to $p(x, y)$. Notice, that $p(x, y)$ isn't exactly a density because it doesn't integrate to 1. $\int P(x, dy) = 1 = \int p(x, y)dy + r(x)$; $\int p(x, y)dy = 1 - r(x)$.

Definition 3. A transition kernel is *reversible* if $\pi(x)p(x, y) = \pi(y)p(y, x)$

Theorem 4. *If a transition kernel is reversible, then π is invariant.*

Proof. We need to check that the definition of invariant distribution is satisfied

$$\begin{aligned}\int_{\Re} \pi(x)P(x,A)dx &= \int_{\Re}\left(\int_A p(x,y)dy\right)\pi(x)dx + \int_{\Re} r(x)\Delta_x(A)\pi(x)dx \\ &= \int_A\int_{\Re} p(x,y)\pi(x)dxdy + \int_A r(x)\pi(x)dx \\ &= \int_A\int_{\Re} p(y,x)\pi(y)dxdy + \int_A r(x)\pi(x)dx \\ &= \int_A \pi(y)\left(\int_{\Re} p(y,x)dx\right)dy + \int_A r(x)\pi(x)dx \\ &= \int_A \pi(y)(1-r(y))dy + \int_A r(x)\pi(x)dx = \pi(A)\end{aligned}$$

□

Metropolis-Hastings

The goal: we want to simulate a draw from the distribution π which we know up to a constant. That is, we can compute a function proportional to π, $f(x) = k\pi(x)$. We will generate a Markov chain with transition kernel of the form (*), that will be reversible for π. Then if the chain will run long enough the element of the chain will have distribution π. The main question is how to generate such a Markov chain?

Suppose we have a Markov chain in state x. Assume that we can draw $y \sim q(x,y)$, a pdf with respect to y (so $\int q(x,y)dy = 1$). Consider using this q as a transition kernel. Notice that if

$$\pi(x)q(x,y) > \pi(y)q(y,x)$$

then the chain won't be reversible (we would move from x to y too often). This suggests that rather than always moving to the new y we draw, we should only move with some probability, $\alpha(x,y)$. If we construct $\alpha(x,y)$ such that

$$\pi(x)q(x,y)\alpha(x,y) = \pi(y)q(y,x)\alpha(y,x)$$

then we will have a reversible transition kernel with invariant measure π. We can take:

$$\alpha(x,y) = \min\{1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\}$$

We can calculate $\alpha(x,y)$ because although we do not know $\pi(x)$, we do know $f(x) = k\pi(x)$, so we can compute the ratio.

In summary, the **Metropolis-Hastings** algorithm is: given x_t we move to x_{t+1} by

1. Generate a draw, y, from $q(x_t, \cdot)$
2. Calculate $\alpha(x_t, y)$
3. Draw $u \sim U[0,1]$
4. If $u < \alpha(x_t, y)$, then $x_{t+1} = y$. Otherwise $x_{t+1} = x_t$

This produces a chain with

$$P(x,dy) = q(y,x)\alpha(y,x)dy + r(x)\Delta_x(dy), \quad r(x) = 1 - \int q(y,x)\alpha(y,x)dy.$$

Then the marginal distribution of x_t will converge to π. In practice, we begin the chain at an arbitrary x_0, run the algorithm many, say M times, then use the last $N < M$ draws as a sample from π. Note that although the marginal distribution of the x_t is π, the x_t are autocorrelated. This is not a problem for computing moments from the draws (although the higher the autocorrelation, the more draws we need to get the same accuracy), but if we want to put standard errors on these moments, we need to take the autocorrelation into account.

Choice of $q()$

- **Random walk chain**: $q(x,y) = q_1(y-x)$, *i.e.* $y = x + \epsilon$, $\epsilon \sim q_1$. This can be a nice choice because if q_1 is symmetric, $q_1(z) = q_1(-z)$, then $\frac{q(x,y)}{q(y,x)}$ drops out of $\alpha(x,y) = \min\{1, \frac{\pi(y)}{\pi(x)}\}$. Popular such q_1 are normal and $U[-a,a]$. Note that there is a tradeoff between step-size in the chain and rejection probability when choosing $\sigma^2 = E\epsilon^2$. Choosing σ^2 too large will lead to many draws of y from low probability areas (low π), and as a result we will reject lots of draws. Choosing σ^2 too small will lead us to accept most draws, but not move very much, and we will have difficulty covering the whole support of π. In either case, the autocorrelation in our draws will be very high and we'll need more draws to get a good sample from π.
- **Independence chain**: $q(x,y) = q_1(y)$
- If there is an additional information that $\pi(y) \propto \psi(y)h(y)$ where ψ is bounded and we can sample from $q(x,y) = h(y)$. This also simplifies $\alpha(x,y) = \min\{1, \frac{\psi(y)}{\psi(x)}\}$
- Autocorrelated $y = a + B(x-a) + \epsilon$ with $B < 0$, this leads to negative autocorrelation in y. The hope is that this reverses some of the positive autocorrelation inherent in the procedure.

14.384 Time Series Analysis

Fall 2013