Markov Chain Monte Carlo Methods

Jesús Fernández-Villaverde
University of Pennsylvania

"Bayesianism has obviously come a long way. It used to be that could tell a Bayesian by his tendency to hold meetings in isolated parts of Spain and his obsession with coherence, self-interrogations, and other manifestations of paranoia. Things have changed..."

Peter Clifford, 1993

Our Goal

- We have a distribution:

$$X \sim f(X)$$

  such that $f > 0$ and $\int f(x)dx < \infty$.

- How do we draw from it?

- We could use Important Sampling...

- ...but we need to find a good source density.

Five Problems

1. A Multinomial Probit Model.

2. A Markov-Switching Model

3. A Stochastic Volatility Model.

4. A Drifting-Parameters VAR Model.

5. A DSGE Model.

A Multinomial Probit Model (MNP)

- MNP goes back to Thurstone (1927) and Bock and Jones (1968).
- An individual i gets utility U_{ij} from choice j, $j \in \{0, 1, ..., J\}$.
- Utility is given by $U_{ij} = x_{ij}\beta + \varepsilon_{ij}$ where ε_{ij} are multivariate normal.
- Examples: car demand, educational choice, voting,...

Problem with MNP

- Under utility maximization, the individual will choose j with probability:

$$
\begin{aligned}
& P\left(U_{ij} > U_{ik}, \text{ for all } k \neq j\right) \\
= & \int_{-\infty}^{\infty} \int_{-\infty}^{U_{ij}} \cdots \int_{-\infty}^{U_{ij}} f\left(U_{i1,\ldots,} U_{iJ}\right) dU_{i1,}\ldots dU_{iJ}
\end{aligned}
$$

where f is the $J-$dimensional normal density.

- Two problems:

1. We need to evaluate a multidimensional normal integral.

2. Conditional on an evaluation of the integral, we need to draw from the posterior or maximize the likelihood.

First Problem: Multidimensional Integral

- Lerman and Manski (1981): Acceptance Sampling.

- GHK (Geweke-Hajivassiliou-Keane) simulator.

Second Problem: Manipulating the Likelihood

- Do we have good importance sampling densities to do so?

- Relation with MSM (McFadden, 1989).

Markov-Switching Model

- Hamilton (1979), Kim and Nelson (1999).

- Regression:

$$z_t = \rho_{s_t} z_{t-1} + e^{\sigma_{s_t}} \varepsilon_t \text{ where } \varepsilon_t \sim \mathcal{N}(0,1)$$

where

$$\begin{aligned} \rho_{s_t} &= \rho_0 S_t + \rho_1 (1 - S_t) \\ \sigma_{s_t} &= \sigma_0 S_t + \sigma_1 (1 - S_t) \end{aligned}$$

and transition matrix for $S_t = \{0,1\}$

$$\begin{pmatrix} \theta & 1-\theta \\ 1-\lambda & \lambda \end{pmatrix}$$

Stochastic Volatility Model

- Changing volatility clustered over time: Kim, Shephard, and Chib (1997).

- We have an autoregressive process:

$$z_t = \rho z_{t-1} + e^{\sigma_t}\varepsilon_t \text{ where } \varepsilon_t \sim \mathcal{N}(0,1)$$

1. and

$$\sigma_t = (1-\lambda)\,\sigma_{mean} + \lambda\sigma_{t-1} + \tau\eta_t \text{ where } \eta_t \sim \mathcal{N}(0,1)$$

- How do we write the likelihood? Comparison with GARCH(p,q) (Engle, 1982, and Bollerslev, 1986).

Drifting-Parameters VAR

- We have a VAR of the form:

$$Y_t = B_t Y_{t-1} + \varepsilon_t \text{ where } \varepsilon_t \sim \mathcal{N}(0, \Sigma)$$

- The parameters B_t drift over time:

$$B_t = B_{t-1} + \omega_t \text{ where } \omega_t \sim \mathcal{N}(0, V)$$

- Cogley and Sargent (2001) and (2002): inflation dynamics in the U.S.

DGSE Models

- We have a likelihood $f\left(Y^T|\theta\right)$ that does not belong to any known parametric family.

- In fact, usually we cannot even write it: only obtain a (possibly stochastic) evaluation.

- Example: basic RBC model.

Transition Kernels I

- The function $P(x, A)$ is a transition kernel for $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$ (a Borel $\sigma-$field on $\mathcal{X}$) such that:

  1. For all $x \in \mathcal{X}$, $P(x, \cdot)$ is a probability measure.

  2. For all $A \in \mathcal{B}(\mathcal{X})$, $P(\cdot, A)$ is measurable.

- When $\mathcal{X}$ is discrete, the kernel is a transition matrix with elements:

$$P_{xy} = P(X_n = y | X_{n-1} = x) \;\; x, y \in \mathcal{X}$$

- When $\mathcal{X}$ is continuous, the kernel is also the conditional density:

$$P(X \in A | x) = \int_A P\left(x, x'\right) dx'$$

Transition Kernels II

- Clearly:

$$P(x, \mathcal{X}) = 1$$

- Also, we allow:

$$P(x, \mathcal{X}) \neq 0$$

- Examples in economics: capital accumulation, job search, prices in financial market,...

Transition Kernels III

Define:

$$P(x, dy) = p(x, y)\, dy + r(x)\, \delta_x(dy)$$

where

1. $p(x, y) \geq 0$, $p(x, x) = 0$

2. $\delta_x(dy)$ is the dirac function in dy,

3. $P(x, x)$, the probability that the chain remains at x, is:

$$r(x) = 1 - \int_{\mathcal{X}} p(x, y)\, dy$$

Markov Chain

- Given a transition kernel P, a sequence $X_0, X_1, ..., X_n, ...$ of random variables is a Markov Chain, denoted by (X_n), if for any t

$$P\left(X_{k+1} \in A | x_0, ..., x_k\right) = P\left(X_{k+1} \in A | x_k\right) = \int_A P\left(x_k, dx\right)$$

- We will only deal with time homogeneous chains, i.e., the distribution of $\left(X_{t_1}, ..., X_{t_k}\right)$ given x_0 is the same as the distribution of $\left(X_{t_1 - t_0}, ..., X_{t_k - t_0}\right)$ given x_0 for every k and every $(k+1)-$uplet $t_0 \leq ... \leq t_k$.

Chapman-Kolmogorov Equations

- For every $(m, n) \in \aleph^2$, $x \in \mathcal{X}$, $A \in \mathcal{B}(\mathcal{X})$

$$P^{m+n}(x, A) = \int_{\mathcal{X}} P^n(y, A) P^m(x, dy)$$

- When $\mathcal{X}$ is discrete, the previous equation is just a matrix product.

- When $\mathcal{X}$ is continuous, the kernel is interpreted as an operator on the space of integrable functions:

$$Ph(x) = \int_A h(y) P(x, dy)$$

  Then, we have a convolution formula: $P^{m+n} = P^m \star P^n$.

Importance of Result

- More in general, we have an operator

$$P\pi(B) = \int_A P(x, B)\,\pi(dx)\,, \text{ for all } A \in \mathcal{B}(\mathcal{X})$$

  where π is a probability distribution.

- We can search for a fixed point:

$$\pi_s = P\pi_s$$

- We say that the distribution π_s is invariant for the transition kernel $P(\cdot, \cdot)$.

Relevant Questions

- Why do we care about a fixed point of the operator?

- Does it exist an invariant distribution?

- Do we converge to it?

- Meyn, S.P. and R.L. Tweedie (1993), *Markov Chains and Stochastic Stability*. Springer-Verlag.

Markov Chain Monte Carlo Methods

- A Markov Chain Monte Carlo ($McMc$) method for the simulation of $f(x)$ is any method producing an ergodic Markov Chain whose invariant distribution is $f(x)$.

- Looking for a Markovian Chain, such that if $X^1, X^2, ..., X^t$ is a realization from it

$$X^t \to X \sim f(x)$$

as t goes to infinity.

Turning the Theory Around

- Note twist we are giving to theory.

- Computing Equilibrium models: we know transition Kernel (from policy functions of the agents) and we compute the invariant distribution.

- McMc: we know invariant distribution and we search for transition kernel that induces that invariant distribution.

- How do we find the transition kernel?

A Trivial Example

- Imagine we want to draw from a binomial with parameter 0.5.

- The simplest way: draw a $u \sim U[0,1]$. If $u \leq 0.5$, then $x = 1$, otherwise $x = 0$.

- The Markov Chain way:

  1. Simulate from transition matrix
  $$\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$$
  with initial state 1.

  2. Every time the state is 1, make $x_t = 1$. Otherwise $x = 0$.

Roadmap

We search for a transition kernel that:

1. Induces an unique stationary distribution with density $f(x)$.

2. Stays within stationary distribution.

3. Converges to the stationary distribution.

4. A Law of Large Number Applies.

5. A Central Limit Theorem Applies.

Searching for a Transition Kernel $P(x, A)$

- Remember that $P(x, dy) = p(x, y)\, dy + r(x)\, \delta_x(dy)$.

- Let $f(x) : \mathcal{X} \to R^+$ be a density.

- Theorem: If $f(x)\, p(x, y) = f(y)\, p(y, x)$, then

$$\int_A f(y)\, dy = \int_{\mathcal{X}} P(x, A)\, f(x)\, dx$$

Proof

$$
\begin{aligned}
&\int_{\mathcal{X}} P(x,A)\, f(x)\, dx \\
&= \int_{\mathcal{X}} \left[\int_A p(x,y)\, dy\right] f(x)\, dx + \int_{\mathcal{X}} r(x)\, \delta_x(A)\, f(x)\, dx = \\
&= \int_A \left[\int_{\mathcal{X}} p(x,y)\, f(x)\, dx\right] dy + \int_A r(x)\, f(x)\, dx = \\
&= \int_A \left[\int_{\mathcal{X}} p(y,x)\, f(y)\, dx\right] dy + \int_A r(x)\, f(x)\, dx = \\
&= \int_A (1 - r(y))\, f(y)\, dy + \int_A r(x)\, f(x)\, dx = \\
&= \int_A f(y)\, dy
\end{aligned}
$$

Remarks

- Note that $\int_A f(y)\,dy = \int_{\mathcal{X}} P(x, A)\, f(x)\, dx$ is an expression for the invariant distribution. We will call that distribution π_s.

- Explanation: if $p(x, y)$ is time reversible, then f is the invariant distribution of $P(x, \cdot)$.

- Time reversibility is the key element we will search for in our McMc algorithms.

Convergence

- Note we have proved that the transition Kernel is a fixed point on the space of densities.

- Can we prove convergence to that invariant distribution?

- If $\{P^n(x,A)\}_{n=0}^{m}$ where $P^n(x,A) = \int_{\mathcal{X}} P(y,A)\, P^{n-1}(x,dy)$ and $P^0(x,A) = P(x,A)$, when do we have that:

$$P^m(x,A) \to \pi_s(A)$$

for π_s-almost all $x \in \mathcal{X}$ as $m \to \infty$ in the total variance distance?

Sufficient Conditions for Convergence

If $P(x, A)$ is such that (1) holds, then the following two conditions about $P(x, A)$ are sufficient for $\Phi^m(x, A) \to \pi_s(A)$ (Smith and Roberts, 1993):

- Irreducibility: if $x \in$support(f)and $A \in \mathcal{B}(\mathcal{X})$, it should be possible to get from x to A with positive probability in a finite number of steps.

- Aperiodicity: The Chain should not have periodic behavior.

Transient period ("burn-in") in our simulations.

A Law of Large Numbers

If $P(x, A)$ is irreducible with invariant distribution π_s, then:

1. π_s is unique.

2. For all π_s-integrable real-valued functions:

$$\frac{1}{M}\sum_{i=1}^{M} h(x_i) \rightarrow \int_{\mathcal{X}} h(x)\,\pi_s(dx)$$

   or

$$\widehat{h} \rightarrow Eh$$

   almost surely.

How do we use this result?

A Central Limit Theorem

- A Central Limit Theorem is useful to study sample-path averages.

- Two conditions on $P(x, A)$:

  1. Positive Harris-Recurrent.

  2. Geometrically Ergodic.

Harris-Recurrence

- A set A is Harris-recurrent if $P_x(\eta_A = \infty) = 1$ for all $x \in A$.

- A Markov Chain is Harris-recurrent if it has an irreducible measure ψ such that for every set A such that $\psi(A) > 0$, A is Harris-recurrent.

- Interpretation (Chan and Geyer, 1994): "Harris recurrence essentially says that there is no measure-theoretic pathology...The main point about Harris recurrence is that asymptotics do not depend on the starting distribution..."

Geometric Ergodicity

- An ergodic Markov chain with invariant distribution π_s is geometrically ergodic if there exist a non-negative real-valued functions bounded in expectation under π_s and a positive constant $r < 1$ such that:

$$\left\| P^M (x, A) - \pi_s (A) \right\| \leq C (x) r^n$$

  for all x and all n and sets A.

- Geometric ergodicity ensures that the distance between the distribution we have and the invariant distribution decreases sufficiently fast.

Chan and Geyer (1994)

If an ergodic Markov chain with invariant distribution π_s is geometrically ergodic, then for all L^2 measurable functions h and any initial distribution

$$M^{0.5}\left(\widehat{h} - Eh\right) \to N\left(0, \sigma_h^2\right)$$

in probability, where:

$$\sigma_h^2 = var\left(h\left(P^0(x, A)\right)\right) + 2\sum_{k=1}^{\infty} cov\left\{h\left(P^0(x, A)\right) h\left(P^0(x, A)\right)\right\}$$

Note the covariance induced by the Markov Chain structure of our problem.

Building our McMc

Previous arguments show that we need to find a transition Kernel $P(x, A)$ such that:

1. It is time reversible.
2. It is irreducible.
3. It is aperiodic.
4. (Bonus Points) It is Harris-recurrent and Geometrically Ergodic.

Note: 1)-4) are sufficient conditions!

McMc and Metropolis-Hastings

- The Metropolis-Hastings algorithm is the ONLY known method of McMc.

- Gibbs-Sampler is a particular form of Metropolis-Hastings.

- Many researchers have proposed almost-but-not-quite-so McMc. Beware of them!.

- Where is the frontier? Perfect Sampling.

On the Use of McMc

- We motivated McMc by the need to draw from a posterior distribution of parameters.

- Up to a point the motivation is misleading.

- Why?

1. McMc helps to draw from a distribution. It does not need to be a posterior. Think of the multivariate integral in the MNP model.

2. McMc explores a distribution. It can be used for classical estimation.

Difficult Problems for Classical Estimation

1. Censored Median Regression for Linear and Non-linear problems (Powell, 1994).

2. Nonlinear IV estimation (Berry, Levinsohn, and Pakes, 1995).

3. Instrumental Quantile Regression.

4. Continuous-updating GMM (Hansen, Heaton, and Yaron, 1996).

5. DSGE Models.

McMc and Classical Estimation I

- Emphasized by Victor Chernozhukov and Han Hong (2003).

- Idea: Laplace-Type Estimators (LTE).

- Define similarly to Bayesian but use general statistical criterion function instead of the likelihood.

- Function $L_n(\theta)$ such that:

$$n^{-1}L_n(\theta) \to M(\theta)$$

McMc and Classical Estimation II

- Define the transformation:

$$p_n(\theta) = \frac{e^{L_n(\theta)}\pi(\theta)}{\int e^{L_n(\theta)}\pi(\theta)\,d\theta}$$

that induces a proper distribution.

- Then, the quasi-posterior mean is:

$$\widehat{\theta} = \int \theta p_n(\theta)\,d\theta$$

can be approximated by draws from a McMc:

$$\widehat{\theta} = \frac{1}{M}\sum_{i=1}^{M}\theta_i$$