Bayesian Estimation Introduction

Paul Schrimpf

UBC
Economics 567

April 1, 2020

References 1

- Brief introductions
  - Mikusheva and Schrimpf (2007) lectures 23-26 (starting slides based off of)
  - Geweke (1999), Geyer (2011)
- Textbooks
  - Widely recommended: Gelman et al. (2013)
  - Econometrics focused: Geweke (2005), Lancaster (2004), Greenberg (2012)
  - Computational Brooks et al. (2011), Marin and Robert (2007), Bolstad (2011)
- Bayesian estimation in IO
  - Jiang, Manchanda, and Rossi (2009): BLP
  - Imai, Jain, and Ching (2009): dynamic discrete choice
  - Gallant, Hong, and Khwaja (2012): dynamic game
  - Norets and Tang (2013): dynamic binary choice
  - Dubé, Hitsch, and Rossi (2010): consumer inertia

Section 1

Introduction

Bayesian econometrics

- Bayesian econometrics is based on two pieces:
  1. A parametric model, giving a distribution, $f(\mathcal{Y}_T|\theta)$, for the data given parameters
  2. A prior distribution for the parameters, $p(\theta)$
- Implies
  - Joint distribution of the data and parameters
    $$p(\mathcal{Y}_T, \theta) = f(\mathcal{Y}_T|\theta)p(\theta)$$
  - Marginal distribution of the data
    $$p(\mathcal{Y}_T) = \int f(\mathcal{Y}_T|\theta)p(\theta)d\theta$$
  - Posterior distribution of parameters
    $$p(\theta|\mathcal{Y}_T) = \frac{f(\mathcal{Y}_T|\theta)p(\theta)}{p(\mathcal{Y}_T)}$$
- Inference based on posterior
  - Report posterior mean (or mode or median) as point estimate
  - Credible set = set of posterior measure $1-\alpha$

Differences between Bayesian and Frequentist Approaches

Frequentist

- θ fixed
- Sample random
- Uncertainty from sampling
- Probability about sampling uncertainty

Bayesian

- θ random
- Sample fixed once observed
- Uncertainty from beliefs about parameter
- Probability about parameter uncertainty

Reasons to be Bayesian

1. Philosophical
2. Bayesian methods asymptotically valid from frequentist perspective
3. Decision theory—leads to admissible decision rules
4. Nuisance parameters easily integrated out
5. Sometimes easier to implement (main reason for this course)

OLS 1

- Model $y_t = x_t\theta + u_t$, $u_t \sim iidN(0,1)$.
- Distribution of data

$$f(Y|X,\theta) = (2\pi)^{-T/2} \exp\left(-\frac{1}{2}(Y - X\theta)'(Y - X\theta)\right)$$

- Conjugate prior (posterior & prior in same family)
  - $\theta \sim N(0, \tau^2 I_k)$,

$$p(\theta) = (2\pi\tau^2)^{-k/2} \exp\left(\frac{-1}{2\tau}\theta'\theta\right)$$

OLS 2

- Posterior

$$p(\theta|Y,X) \propto \exp\left(-\frac{1}{2}\left[-Y'X\theta - \theta'X'Y + \theta'X'X\theta + \frac{1}{\tau^2}\theta'\theta\right]\right)$$
$$\propto \exp\left(-\frac{1}{2}\left[-Y'X\theta - \theta'X'Y + \theta'(X'X + \frac{I_k}{\tau^2})\theta\right]\right)$$
$$\propto \exp\left(-\frac{1}{2}\left[\left(\theta - (X'X + \frac{I_k}{\tau^2})^{-1}X'Y\right)'(X'X + \frac{I_k}{\tau^2})^{-1}\left(\theta - (X'X + \frac{I_k}{\tau^2}\right.\right.\right.$$

so $\theta|Y,X \sim N(\tilde{\theta},\tilde{\Sigma})$ with

$$\tilde{\theta} = (X'X + \frac{I_k}{\tau^2})^{-1}X'Y$$
$$\tilde{\Sigma} = (X'X + \frac{I_k}{\tau^2})^{-1}$$

- Fix τ and $T \to \infty$ with $\frac{X'X}{T} \to Q_{XX}$, then $\tilde{\theta} \to \theta_0$
- Uninformative prior, $\tau \to \infty$, $\tilde{\theta} \to (X'X)^{-1}X'Y = \hat{\theta}^{ML}$

Section 2

MCMC

MCMC

- Posterior

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

- Closed form posterior is rare, often impossible
- Sample $\theta_i \sim p(\theta|\mathcal{Y}_T)$ instead
- Markov Chain Monte-Carlo

Acceptance-Rejection 1

- Want $\xi \sim \pi(x)$, can calculate $f(x) \propto \pi(x)$
- Find distribution with pdf $h(x)$ such that $f(x) \leq ch(x)$
- Accept-reject
  1. Draw $z \sim h(x)$, $u \sim U[0,1]$
  2. If $u \leq \frac{f(z)}{ch(z)}$, then $\xi = z$. Otherwise repeat (1)

Acceptance-Rejection 2

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

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

- Advantage: directly gives independent draws from π
- Downside: if h too far from f, then will reject many draws

Markov Chains 1

- Transition kernel $P(x, A)$ = probability of moving from x into the set A.
- Distribution of x^k is π^*, then the distribution of $y = x^{k+1}$ is

$$\tilde{\pi}(y)dy = \int_{\Re} \pi^*(x)P(x, dy)dx$$

- Invariant measure if $\tilde{\pi} = \pi^*$
- Invariant measure exists iff:
  - Irreducible: every state can be reached from any other
  - Positive recurrent: $\mathrm{E}[$ time until x again $|x]$ finite
- Conditions for chain to converge to invariant measure from any initial measure:
  - Irreducible
  - Positive recurrent
  - Aperiodic: greatest common denominator of $\{n : y$ can be reached from x in n steps$\}$ is 1

Markov Chains 2

- Easier sufficient condition for convergence:
    - Reversible: if $\pi(x)p(x,y) = \pi(y)p(y,x)$ (aka detailed balance)
- Goal: construct a Markov chain, which we can simulate, that has the posterior as its invariant measure and has fast mixing

Metropolis-Hastings 1

- General purpose method to sample from π
  1. Draw $y \sim q(x^j, \cdot)$
  2. Calculate $\alpha(x^j, y) = \min\{1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\}$
  3. Draw $u \sim U[0, 1]$
  4. If $u < \alpha(x^j, y)$, then $x^{j+1} = y$. Otherwise $x^{j+1} = x^j$
- Invariant measure is π
  - Proof: detailed balance condition

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

- Proposal density q
  - Too disperse $\implies$ many rejections
  - Too concentrated $\implies$ high autocorrelation and slow mixing (slow convergence to π)
  - Common choices:
    - Random walk chain: $q(x, y) = q_1(y - x)$, e.g. $y = x + \epsilon$, $\epsilon \sim N(0, s)$
    - Independence chain: $q(x, y) = q_1(y)$
    - Autocorrelated $y = a + B(x - a) + \epsilon$ with $B < 0$

Gibbs sampling 1

- Break x into blocks $x = (x_1, x_2, \ldots, x_d)$ such that we can draw from

$$\pi(x_k | x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d) \quad \forall k$$

- Simulate
  - $x_1^{(j+1)}$ from $\pi(x_1^{(j+1)} | x_2^{(j)}, \ldots, x_d^{(j)})$
  - $x_2^{(j+1)}$ from $\pi(x_2^{(j+1)} | x_1^{(j+1)}, x_3^{(j)}, \ldots, x_d^{(j)})$
  - $x_3^{(j+1)}$ from $\pi(x_3^{(j+1)} | x_1^{(j+1)}, x_2^{(j+1)}, x_4^{(j)}, \ldots, x_d^{(j)})$
  - ...
- Can be viewed as Metropolis-Hastings with $q = \pi(\cdot|\cdot)$
- Pros:
  - Usually fast
  - No need to choose candidate distribution
  - Sometimes less autocorrelation
  - Easy to incorporate latent variables (data augmentation)

Gibbs sampling 2

- Cons:
  - Not possible for all models & priors
  - Can lead to slow mixing (especially with many blocks)
  - "many naive users still have a preference for Gibbs updates that is entirely unwarranted. If I had a nickel for every time someone had asked for help with slowly converging MCMC and the answer had been to stop using Gibbs, I would be rich. Use Gibbs updates only if the resulting sampler works well. If not, use something else." Geyer (2011)

Gibbs sampling 1

- Example: probit
  - $d = \{x\beta + \epsilon > 0\}$, $\epsilon \sim N(0,1)$
  - Prior: $\beta \sim N(0, I\tau^2)$
  - Posterior: $\prod_i \Phi(x_i\beta)^{d_i} (1 - \Phi(x_i\beta))^{1-d_i}$
  - Data augmentation: draw $y_i = x_i\beta + \epsilon_i$ conditional on data and β
  - Gibbs sampler:
    - Draw $y_i \sim$ truncated $N(x_i\beta, 1; d_i)$
    - Draw $\beta \sim N\left((X'X + \frac{I_k}{\tau^2})^{-1}X'Y, (X'X + \frac{I_k}{\tau^2})^{-1}\right)$
- Example: random coefficients probit (in Bayesian stats, random coefficients $\approx$ multilevel model)
  - $d_{it} = \{x_{it}\beta_i + \epsilon_{it} > 0\}$, $\epsilon_{it} \sim N(0,1)$, $\beta_i \sim N(\beta, \Sigma)$
  - Prior: $\beta \sim N(0, I\tau^2)$, $\Sigma^{-1} \sim$ Wishart (V)
  - Data augmentation: draw $y_{it} = x_{it}\beta_i + \epsilon_{it}$ conditional on data and β_i
  - Gibbs sampler:
    - Draw $y_{it} \sim$ truncated $N(x_{it}\beta_i, 1; d_i)$

Gibbs sampling 2

- Draw $\beta_i \sim N\left((X_i'X_i + \Sigma)^{-1}X'Y, (X_i'X_i + \Sigma)^{-1}\right)$
- Draw $\beta \sim N\left(1/n \sum_i \beta_i, S\right)$
- Draw $\Sigma^{-1} \sim$ Wishart(something)

- Software: OpenBUGS, WinBUGS, JAGS

Hamiltonian MCMC 1

- Improve Metropolis-Hastings through better choice of candidate density
  - Avoid high autocorrelation
- Overview: Neal (2011)
- Software: STAN, Turing.jl, DynamicHMC.jl

Hamiltonian MCMC 1

- Hamiltonian dynamics:
  - Parameters = position = x
  - Momentum = m
  - Hamiltonian $H(x, m) = U(x) + K(m) =$ potential + kinetic energy
  - $U(x) = -\log(\pi(x))$
  - Conservation of energy $\rightarrow$

$$\frac{dx}{dt} = \frac{\partial H}{\partial m}$$
$$\frac{dm}{dt} = -\frac{\partial H}{\partial x}$$

    Given H can accurately compute $x(t)$, $m(t)$
  - Useful properties:
    - Symmetrically invertible: $(x^*, m^*) = T_s(x, m) \iff (x, -m) = T_s(x^*, -m^*)$
    - Conserved: $\frac{dH}{dt} = 0$

Hamiltonian MCMC 2

- Volume preserved: mapping $T_s : (x(t), m(t)) \rightarrow (x(t+s), m(t+s))$ has jacobian, B_s, with determinant 1
- Use Hamiltonian dynamics for candidate density

Hamiltonian MCMC 1

- Hamiltonian MC for drawing from π
  - Set $U(x) = -\log pi(x)$, $K(m) = m^T M^{-1} m/2$
  - Each step of chain:
    1. Draw $m \sim N(0, M)$
    2. Simulate dynamics $(x^*, m^*) = T_s(x, m)$
    3. Accept x^* with probability

$$\alpha(x^*, m^*; x, m) = \min\left\{1, \exp\left(-U(x^*) + U(x) - K(m^*) + K(m)\right)\right\}$$

- Detailed balance:

$$p(x_1; x_0) \propto \begin{cases} 0 & \text{if } (x_1, m_1) \neq T_s(x_0, m_0) \text{ for any } m_0 \\ e^{-{m_0}^T M^{-1} m_0/2} \alpha(x_1, m_1; x_0, m) & \text{if } (x_1, m_1) = T_s(x_0, m_0) \end{cases}$$

$$\propto \begin{cases} 0 & \text{if } (x_1, m_1) \neq T_s(x_0, m_0) \text{ f} \\ e^{-K(m_0)} \min\{1, e^{-U(x_1)+U(x_0)-K(m_1)+K(m_0)}\} & \text{if } (x_1, m_1) = T_s(x_0, m_0) \end{cases}$$

Hamiltonian MCMC 2

$\pi(x) = \exp(-U(x))$, and $(x_1, m_1) = T_s(x_0, m_0)$ implies $(x_0, m_0) = T_s(x_1, -m_1)$, so

$$
\begin{aligned}
\pi(x_0)p(x_1; x_0) =& e^{-U(x_0)-K(m_0)} \min\{1, e^{-U(x_1)+U(x_0)-K(m_1)+K(m_0)}\} \\
=& e^{-U(x_1)-K(-m_1)} \min\{1, e^{-U(x_0)+U(x_1)-K(m_0)+K(-m_1)}\} \\
=& \pi(x_1, m_1)p(x_0, m_0; x_1, m_1)
\end{aligned}
$$

Hamiltonian MCMC 1

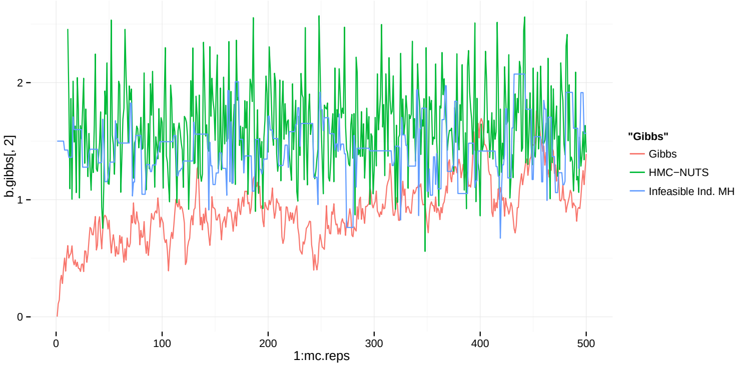
- Tuning choices:
    - Length of path to simulate, s, which in practice is discretized into L steps of size ϵ
    - Variance of momentum, M
    - Various methods to automate e.g. NUTS (used by Stan)

Example: Linear Regression

Julia Code and Notes

Example: Probit

R code



Bolstad, William M. 2011. *Understanding computational Bayesian statistics*, vol. 644. John Wiley & Sons. URL http://onlinelibrary.wiley.com.ezproxy.library.ubc.ca/doi/10.1002/9780470567371.ch10/summary.

Brooks, Steve, Andrew Gelman, Galin Jones, and Xiao-Li Meng. 2011. *Handbook of Markov Chain Monte Carlo*. CRC Press.

Dubé, Jean-Pierre, Günter J. Hitsch, and Peter E. Rossi. 2010. "State dependence and alternative explanations for consumer inertia." *The RAND Journal of Economics* 41 (3):417–445. URL http://dx.doi.org/10.1111/j.1756-2171.2010.00106.x.

Gallant, A Ronald, Han Hong, and Ahmed Khwaja. 2012. "Bayesian Estimation of a Dynamic Game with Endogenous, Partially Observed, Serially Correlated State." Tech. rep., Duke University, Department of Economics. URL http://econpapers.repec.org/paper/dukdukeec/12-01.htm.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian data analysis*. CRC press.

Geweke, John. 1999. "Using simulation methods for bayesian econometric models: inference, development,and communication." *Econometric Reviews* 18 (1):1–73. URL http://dx.doi.org/10.1080/07474939908800428.

———. 2005. *Contemporary Bayesian econometrics and statistics*, vol. 537. John Wiley & Sons.

Geyer, Charles. 2011. "Introduction to Markov Chain Monte Carlo." *Handbook of Markov Chain Monte Carlo* :3–48URL http://www.mcmchandbook.net/HandbookChapter1.pdf.

Greenberg, Edward. 2012. *Introduction to Bayesian econometrics*. Cambridge University Press.

Imai, Susumu, Neelam Jain, and Andrew Ching. 2009. "Bayesian estimation of dynamic discrete choice models." *Econometrica* 77 (6):1865–1899. URL http://onlinelibrary.wiley.com/doi/10.3982/ECTA5658/abstract.

Jiang, R., P. Manchanda, and P.E. Rossi. 2009. "Bayesian analysis of random coefficient logit models using aggregate data." *Journal of Econometrics* 149 (2):136–148. URL http://www.sciencedirect.com/science/article/pii/S0304407608002297.

Lancaster, Tony. 2004. *An introduction to modern Bayesian econometrics*. Blackwell Oxford. URL http://www.econ.brown.edu/faculty/Tony_Lancaster/papers/book_selection.pdf.

Marin, Jean-Michel and Christian Robert. 2007. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics: A Practical Approach to Computational Bayesian Statistics*. Springer. URL http://link.springer.com.ezproxy.library.ubc.ca/book/10.1007%2F978-0-387-38983-7.

Mikusheva, Anna and Paul Schrimpf. 2007. "14.384 Time Series Analysis, Fall 2007 (revised 2009)." URL http://ocw.mit.edu/courses/economics/14-384-time-series-analysis-fall-2013/lecture-notes/.

Neal, Radford. 2011. "MCMC using Hamiltonian dynamics." *Handbook of Markov Chain Monte Carlo* 2. URL http://arxiv.org/abs/1206.1901.

Norets, A. and X. Tang. 2013. "Semiparametric Inference in Dynamic Binary Choice Models." *The Review of Economic Studies* URL http://restud.oxfordjournals.org/content/early/2013/12/15/restud.rdt050.abstract.