Introduction to Bayesian Estimation

June 9, 2011

The Plan

1. Talk about Bayesian methods
2. Estimate a simple AR(1) using Bayesian methods
   2.1 Estimate posterior by MCMC methods (Random Walk Metroplis Algorithm)
3. Constructing probability intervals for model outputs

Most of today's lecture can be found in Eddy (2004), An and Schorfheide (2007) and in Canova's textbook *Methods for Applied Macroeconomics*

Frequentist vs Bayesian statistics

Variance of estimator vs variance of parameter

- Parameters have distributions, just like the shocks
- Variance of estimator (frequentist) vs variance of parameter (Bayesian)

Frequentist confidence intervals:

- If point estimates are the truth and with repeated draws from the population of equal sample length, what is the interval that $\widehat{\theta}$ lies in 95 per cent of the time?

Bayesian probability intervals:

- Conditional on the observed data, a prior distribution of θ and a functional form (i.e. model), what is the interval that high the highest probability of including θ?

The main conceptual difference is how the data and parameters are treated, not the use of priors

Bayesian statistics

Bayesians used to be fringe types, but are now more or less the mainstream in macro,

- This is largely due to increased computing power

Bayesian methods have several advantages:

- Facilitates incorporating information from outside of sample
- Easy to compute confidence/probabilty intervals of functions of parameters

It is arguably the only approach that can be made consistent within a decision theoretic framework

What is Bayesian statistics?

Subjective view of probability

- What does subjective mean?
- Coin flip: What is prob(heads) after flip but before looking?
- Probabilities can be viewed as statements about our knowledge of a parameter, even if the parameter is constant

Is subjectiveness good?

- Yes, it allow people to incorporate information from outside the sample
- Yes, because it is always there (it just more hidden in frequentist methods)
- Yes, because one can use "objective", or non-informative priors
- And: Sensitivity to priors can be checked

From last time

We can find the maximum likelihood estimate of a parameter vector Θ from the state space system

$$
\begin{aligned}
X_t &= AX_{t-1} + Cu_t \\
Z_t &= DX_t + v_t
\end{aligned}
$$

where A, C, D and $E(v_t v_t')$ are functions of θ.

- We are normally interested in more than point estimates: For instance, how confident are we about the estimated parameters?

Inference with ML

One approach that works well if likelihood function is well behaved (i.e. asymptotically...) is to compute the Hessian (i.e. the matrix of second derivatives) of the likelihood function. The inverse of the negative of the Hessian gives an estimate of the covariance matrix of the parameter vector θ, i.e.

$$E\left(\widehat{\theta}-\theta\right)\left(\widehat{\theta}-\theta\right)' \simeq \left[-\frac{\partial^2 \mathcal{L}(Z \mid \theta)}{\partial\theta\partial\theta'}\mid_{\theta=\widehat{\theta}}\right]^{-1}$$

- In practice, this approach can be quite sensitive to step length when taking numerical derivatives

Concepts and notation

The parts we need to make Bayesian inference are:

- Data (observables) Z
- (Unobservable) parameters θ
- A prior $p(\theta)$
- A model which make joint probability statements about the observables and the parameters $p(Z \mid \theta)$

Bayesian statistics

Most of Bayesian econometrics consists of simulating distributions of parameters using numerical methods.

- A simulated posterior is a numerical approximation to the likelihood function (combined with prior density)
- We want to construct a posterior estimate of the distribution $p(Z \mid \theta)p(\theta)$ since by Bayes' rule this is proportional to $p(\theta \mid Z)$

The most popular procedure to simulate the posterior is called the Random Walk Metropolis Algorithm

The Random-Walk Metropolis Algorithm

1. Start with an arbitrary value θ_0
2. Update from θ_j to $\theta_{j+1} (j = 1, 2, \ldots J)$ by
   2.1 Generate $\theta^* \sim N(\theta_j, \Sigma)$
   2.2 Define
   $$\alpha = \min\left(\frac{L(Y \mid \theta^*)}{L(Y \mid \theta_j)}, 1\right) \tag{1}$$
   2.3 Take
   $$\theta_{j+1} = \left\{ \begin{array}{c} \theta^* \text{ with probability } \alpha \\ \theta_j \text{ otherwise} \end{array} \right\}$$
3. Repeat Step 2 J times

The Random-Walk Metropolis Algorithm

What do we need for the algorithm?

1. A starting value for θ_j
2. A covariance matrix for the random walk component that generates the candidate, or proposal draw θ_j^*
3. We must know how to evaluate the likelihood function

Practical issues

How do we choose starting value $\theta^{(0)}$?

- Asymptotically this shouldn't matter, but matters greatly in practice
- One (good) option is to first maximize the likelihood and choose the MLE of Θ as $\theta^{(0)}$

How do we choose covariance Σ?

- Again, asymptotically this shouldn't matter, but matters greatly in practice
- Need to scale to achieve optimal acceptance ratio of 23%-40%
    - Scale for each element of θ
    - Scale overall variance

Simple example

Estimate a AR(1) using the Metropolis Algorithm

$$x_t = \rho x_{t-1} + u_t : \Theta = \left\{\rho, \sigma_u^2\right\}$$

We need:

- Data
  - Use simulated data
- A prior
  - (Improper) uniform
- The likelihood function

$$\ln \mathcal{L}(x^t \quad | \quad \Theta) = -.5 \sum_{t=2}^{T} \left[\ln(2\pi) + \ln \left|\sigma_u^2\right| + u_t' \left(\sigma_u^2\right)^{-1} u_t \right]$$

$$-.5 \left[\ln(2\pi) - \ln \left[(1-\rho^2)^{-1}\sigma_u^2\right] - \left(\frac{1-\rho^2}{\sigma_u^2}\right) x_1^2 \right]$$

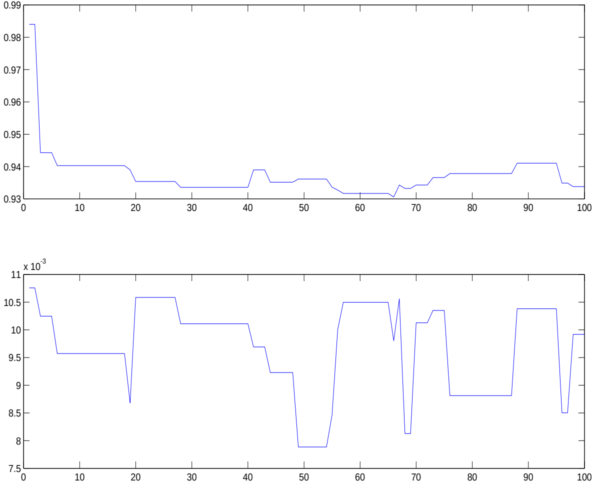
Simple example

Estimate a AR(1) using the Metropolis Algorithm

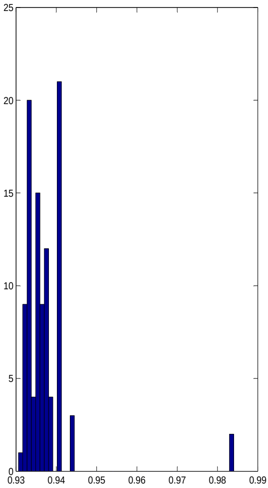
$$x_t = \rho x_{t-1} + u_t : \Theta = \{\rho, \sigma_u^2\}$$

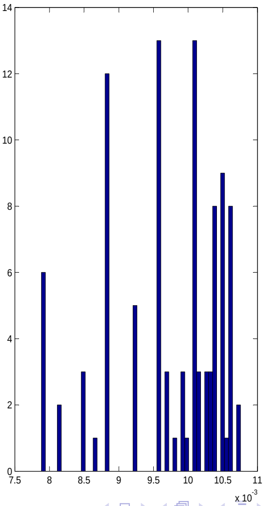
We will set $\rho = 0.9$ and $\sigma_u^2 = 0.01$ and generate an artificial sample with $T = 100$.

The MCMC with J=100

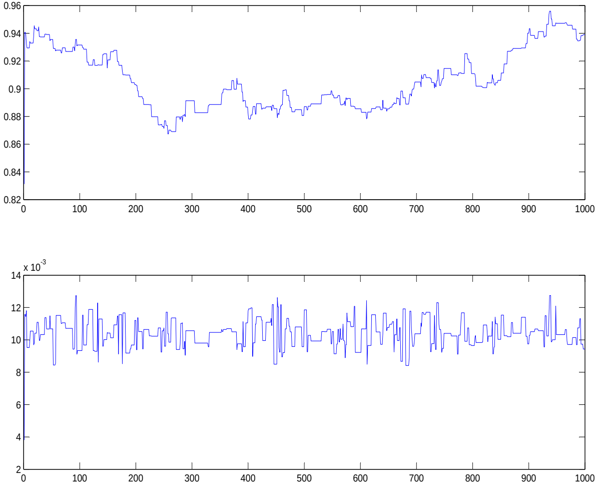


The Histograms of MCMC with J=100

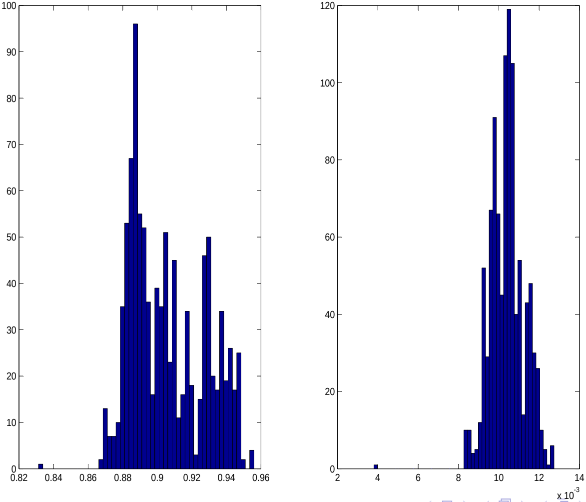




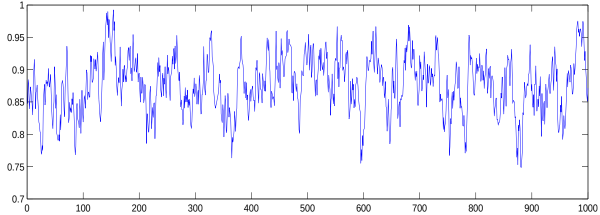
The MCMC with J=1000

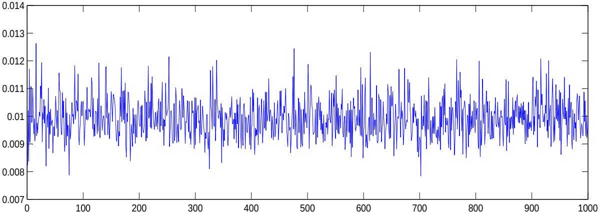


The Histograms of MCMC with J=1000

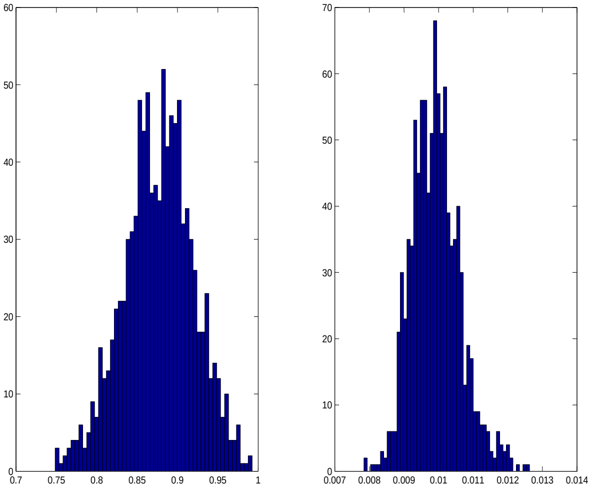


The MCMC with J=1000 000

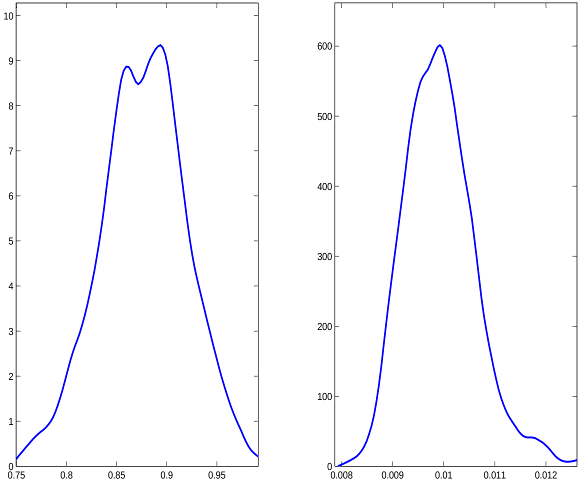




The Histograms of MCMC with J=1000 000



Estimated posterior density with J=1000



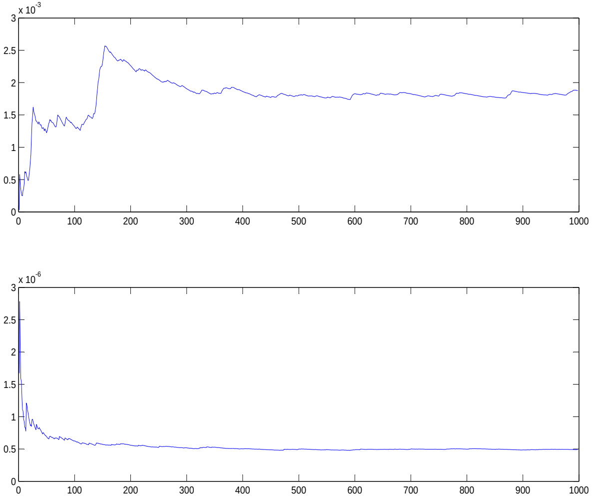
Convergence

How many draws do we need?

- Optimal J increases with number of parameters
- One informal check is to plot the diagonal of the recursive covariance matrix of the MCMC

$$\frac{1}{j}\sum_{i=0}^{j}\theta^{(i)}\theta'^{(i)} \text{ for } j = 1, 2, \dots J$$

Plot of recursive covariance of MCMC with J=1000



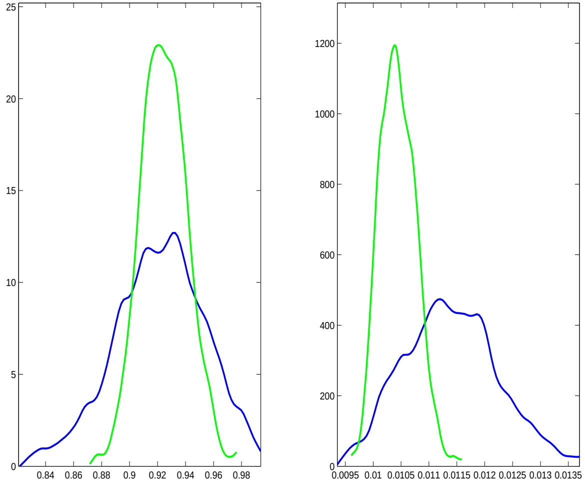
The posterior density of θ

Remember: It describes our knowledge of θ

- We should have more precise knowledge with a longer sample

Let's re-estimate the posterior with T=500;

More data is good!



Probability intervals of functions of θ

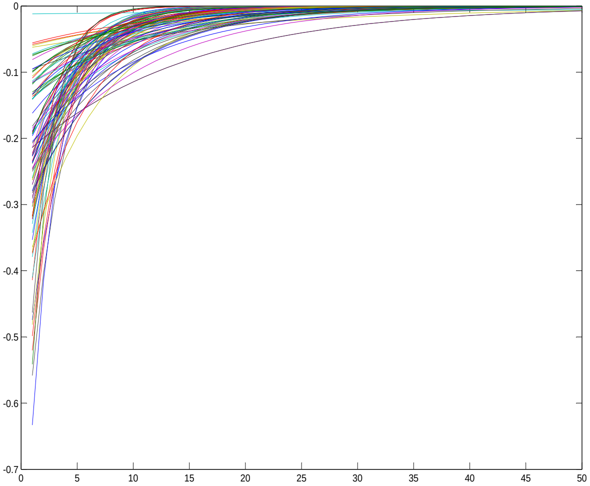
We now know how to find the posterior distribution of the parameter vector θ
We can use the MCMC to also construct probability intervals of any function $h(\theta)$

1. Draw an integer j on a uniform distribution between 1 and J
2. Compute $h(\theta_j)$ and save.
3. Repeat steps 1 and 2 "many" times (but usually fewer than J times are necessary).
4. Find percentiles of the saved outputs from $h(\theta)$. These are the probability intervals of $h(\theta)$.

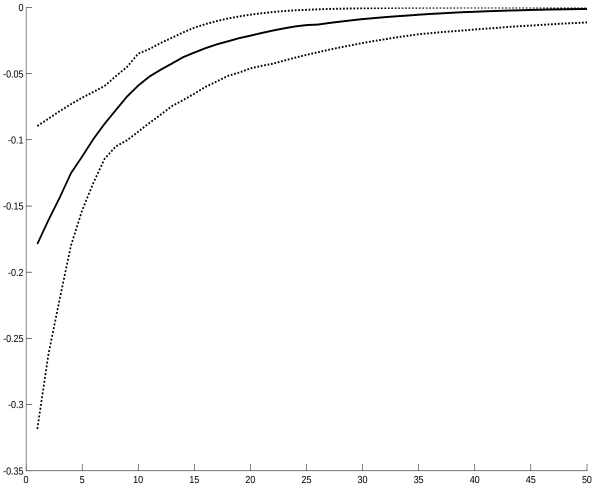
Example: Probability intervals for impulse response function

1. Draw S integers on a uniform distribution between 1 and J
2. Compute $\rho^s \sigma_u$ using θ_j for $j = \{s_1, s_2, ..., s_S\} = \{4, 6738, ..., 435\}$ and save results.
3. Repeat steps 1 and 2 500 times.
4. Find percentiles of the saved outputs from $\rho^s \sigma_u$ using θ_j for each horizon s.
5. Plot.

Unsorted IRF S=100



Mean and 90% prob interval of IRF to u_t^x



That's it for today.