Bayesian Inference for DSGE Models

Lawrence J. Christiano

Outline

- State space-observer form.
  - convenient for model estimation and many other things.
- Bayesian inference
  - Bayes' rule.
  - Monte Carlo integation.
  - MCMC algorithm.
  - Laplace approximation

State Space/Observer Form

- Compact summary of the model, and of the mapping between the model and data used in the analysis.
- Typically, data are available in log form. So, the following is useful:
  - If x is steady state of x_t :

$$\begin{aligned} \hat{x}_t &\equiv \frac{x_t - x}{x}, \\ &\Longrightarrow \frac{x_t}{x} = 1 + \hat{x}_t \\ &\Longrightarrow \log\left(\frac{x_t}{x}\right) = \log\left(1 + \hat{x}_t\right) \approx \hat{x}_t \end{aligned}$$

- Suppose we have a model solution in hand:[1]

$$\begin{aligned} z_t &= A z_{t-1} + B s_t \\ s_t &= P s_{t-1} + \epsilon_t,\ E\epsilon_t\epsilon_t' = D. \end{aligned}$$

[1] Notation taken from solution lecture notes, http://faculty.wcas.northwestern.edu/~lchrist/course/Korea 2012/lecture on solving rev.pdf

State Space/Observer Form

- Suppose we have a model in which the date t endogenous variables are capital, K_{t+1}, and labor, N_t:

$$z_t = \begin{pmatrix} \hat{K}_{t+1} \\ \hat{N}_t \end{pmatrix},\ s_t = \hat{\varepsilon}_t,\ \epsilon_t = e_t.$$

- Data may include variables in z_t and/or other variables.
  - for example, suppose available data are N_t and GDP, y_t and production function in model is:

$$y_t = \varepsilon_t K_t^{\alpha} N_t^{1-\alpha},$$

    so that

$$\begin{aligned} \hat{y}_t &= \hat{\varepsilon}_t + \alpha \hat{K}_t + (1-\alpha)\hat{N}_t \\ &= \begin{pmatrix} 0 & 1-\alpha \end{pmatrix} z_t + \begin{pmatrix} \alpha & 0 \end{pmatrix} z_{t-1} + s_t \end{aligned}$$

- From the properties of $\hat{y}_t$ and $\hat{N}_t$:

$$Y_t^{data} = \begin{pmatrix} \log y_t \\ \log N_t \end{pmatrix} = \begin{pmatrix} \log y \\ \log N \end{pmatrix} + \begin{pmatrix} \hat{y}_t \\ \hat{N}_t \end{pmatrix}$$

State Space/Observer Form

- Model prediction for data:

$$
\begin{aligned}
Y_t^{data} &= \begin{pmatrix} \log y \\ \log N \end{pmatrix} + \begin{pmatrix} \hat{y}_t \\ \hat{N}_t \end{pmatrix} \\
&= \begin{pmatrix} \log y \\ \log N \end{pmatrix} + \begin{bmatrix} 0 & 1-\alpha \\ 0 & 1 \end{bmatrix} z_t + \begin{bmatrix} \alpha & 0 \\ 0 & 0 \end{bmatrix} z_{t-1} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} s_t \\
&= a + H\xi_t
\end{aligned}
$$

$$
\xi_t = \begin{pmatrix} z_t \\ z_{t-1} \\ \hat{\varepsilon}_t \end{pmatrix},\ a = \begin{bmatrix} \log y \\ \log N \end{bmatrix},\ H = \begin{bmatrix} 0 & 1-\alpha & \alpha & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}
$$

- The *Observer Equation* may include measurement error, w_t :

$$
Y_t^{data} = a + H\xi_t + w_t,\ Ew_t w_t' = R.
$$

- Semantics: ξ_t is the *state* of the system (do not confuse with the economic state (K_t, ε_t)!).

State Space/Observer Form

- Law of motion of the state, ξ_t (state-space equation):

$$\xi_t = F\xi_{t-1} + u_t,\ Eu_t u_t' = Q$$

$$\begin{pmatrix} z_{t+1} \\ z_t \\ s_{t+1} \end{pmatrix} = \begin{bmatrix} A & 0 & BP \\ I & 0 & 0 \\ 0 & 0 & P \end{bmatrix} \begin{pmatrix} z_t \\ z_{t-1} \\ s_t \end{pmatrix} + \begin{pmatrix} B \\ 0 \\ I \end{pmatrix} \epsilon_{t+1},$$

$$u_t = \begin{pmatrix} B \\ 0 \\ I \end{pmatrix} \epsilon_t,\ Q = \begin{bmatrix} BDB' & 0 & BD \\ 0 & 0 & 0 \\ DB' & & D \end{bmatrix},\ F = \begin{bmatrix} A & 0 & BP \\ I & 0 & 0 \\ 0 & 0 & P \end{bmatrix}.$$

Uses of State Space/Observer Form

- Estimation of θ and forecasting ξ_t and Y_t^{data}
- Can take into account situations in which data represent a mixture of quarterly, monthly, daily observations.
- 'Data Rich' estimation. Could include several data measures (e.g., employment based on surveys of establishments and surveys of households) on a single model concept.
- Useful for solving the following forecasting problems:
  - Filtering (mainly of technical interest in computing likelihood function):

$$P\left[\xi_t | Y_{t-1}^{data}, Y_{t-2}^{data}, ..., Y_1^{data}\right],\ t = 1, 2, ..., T.$$

  - Smoothing:

$$P\left[\xi_t | Y_T^{data}, ..., Y_1^{data}\right],\ t = 1, 2, ..., T.$$

  - Example: 'real rate of interest' and 'output gap' can be recovered from ξ_t using simple New Keynesian model.
- Useful for deriving a model's implications vector autoregressions

Mixed Monthly/Quarterly Observations

- Different data arrive at different frequencies: daily, monthly, quarterly, etc.
- This feature can be easily handled in state space-observer system.
- Example:
  - suppose inflation and hours are monthly, $t = 0$, $1/3$, $2/3$, 1, $4/3$, $5/3$, $2, \ldots$
  - suppose gdp is quarterly, $t = 0, 1, 2, 3, \ldots$.

$$Y_t^{data} = \begin{pmatrix} GDP_t \\ \text{monthly inflation}_t \\ \text{monthly inflation}_{t-1/3} \\ \text{monthly inflation}_{t-2/3} \\ \text{hours}_t \\ \text{hours}_{t-1/3} \\ \text{hours}_{t-2/3} \end{pmatrix}, \ t = 0, 1, 2, \ldots .$$

  that is, we can think of our data set as actually being quarterly, with quarterly observations on the first month's inflation, quarterly observations on the second month's inflation, etc.

Mixed Monthly/Quarterly Observations

- Problem: find state-space observer system in which observed data are:

$$Y_t^{data} = \begin{pmatrix} GDP_t \\ \text{monthly inflation}_t \\ \text{monthly inflation}_{t-1/3} \\ \text{monthly inflation}_{t-2/3} \\ \text{hours}_t \\ \text{hours}_{t-1/3} \\ \text{hours}_{t-2/3} \end{pmatrix},\ t = 0, 1, 2, \ldots .$$

- Solution: easy!

Mixed Monthly/Quarterly Observations

- Model timing: $t = 0, 1/3, 2/3, ...$

$$\begin{aligned} z_t &= Az_{t-1/3} + Bs_t, \\ s_t &= Ps_{t-1/3} + \epsilon_t,\ E\epsilon_t\epsilon_t' = D. \end{aligned}$$

- Monthly state-space observer system, $t = 0, 1/3, 2/3, ...$

$$\xi_t = F\xi_{t-1/3} + u_t,\ Eu_tu_t' = Q,\ u_t\tilde{}iid\ t = 0, 1/3, 2/3, ...$$

$$Y_t = H\xi_t,\ Y_t = \begin{pmatrix} y_t \\ \pi_t \\ h_t \end{pmatrix}.$$

- Note:

$$\overbrace{\xi_t = F^3\xi_{t-1} + u_t + Fu_{t-1/3} + F^2u_{t-2/3}}^{\text{first order vector autoregressive representation for quarterly state}},$$

$$u_t + Fu_{t-1/3} + F^2u_{t-2/3}\ \tilde{}\ \underline{iid\ for\ t = 0, 1, 2, ...}!!$$

Mixed Monthly/Quarterly Observations

Consider the following system:

$$\begin{pmatrix} \xi_t \\ \xi_{t-\frac{1}{3}} \\ \xi_{t-\frac{2}{3}} \end{pmatrix} = \begin{bmatrix} F^3 & 0 & 0 \\ F^2 & 0 & 0 \\ F & 0 & 0 \end{bmatrix} \begin{pmatrix} \xi_{t-1} \\ \xi_{t-\frac{4}{3}} \\ \xi_{t-\frac{5}{3}} \end{pmatrix} + \begin{bmatrix} I & F & F^2 \\ 0 & I & F \\ 0 & 0 & I \end{bmatrix} \begin{pmatrix} u_t \\ u_{t-\frac{1}{3}} \\ u_{t-\frac{2}{3}} \end{pmatrix}.$$

Define

$$\tilde{\xi}_t = \begin{pmatrix} \xi_t \\ \xi_{t-\frac{1}{3}} \\ \xi_{t-\frac{2}{3}} \end{pmatrix}, \tilde{F} = \begin{bmatrix} F^3 & 0 & 0 \\ F^2 & 0 & 0 \\ F & 0 & 0 \end{bmatrix}, \ \tilde{u}_t = \begin{bmatrix} I & F & F^2 \\ 0 & I & F \\ 0 & 0 & I \end{bmatrix} \begin{pmatrix} u_t \\ u_{t-\frac{1}{3}} \\ u_{t-\frac{2}{3}} \end{pmatrix},$$

so that

$$\tilde{\xi}_t = \tilde{F}\tilde{\xi}_{t-1} + \tilde{u}_t, \ \tilde{u}_t \tilde{\ } iid \text{ in quarterly data, } t = 0, 1, 2, ...$$

$$E\tilde{u}_t\tilde{u}_t' = \tilde{Q} = \begin{bmatrix} I & F & F^2 \\ 0 & I & F \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} D & 0 & 0 \\ 0 & D & 0 \\ 0 & 0 & D \end{bmatrix} \begin{bmatrix} I & F & F^2 \\ 0 & I & F \\ 0 & 0 & I \end{bmatrix}'$$

Mixed Monthly/Quarterly Observations

- Conclude: state space-observer system for mixed monthly/quarterly data, for $t = 0, 1, 2, \ldots$

$$\tilde{\xi}_t = \tilde{F}\tilde{\xi}_{t-1} + \tilde{u}_t, \ \tilde{u}_t \tilde{\ } iid, \ E\tilde{u}_t\tilde{u}_t' = \tilde{Q},$$

$$Y_t^{data} = \tilde{H}\tilde{\xi}_t + w_t, \ w_t \tilde{\ } iid, \ Ew_tw_t' = R.$$

- Here, $\tilde{H}$ selects elements of $\tilde{\xi}_t$ needed to construct Y_t^{data}
  - can easily handle distinction between whether quarterly data represent monthly averages (as in flow variables), or point-in-time observations on one month in the quarter (as in stock variables).
- Can use Kalman filter to forecast ('nowcast') current quarter data based on first month's (day's, week's) observations.

Connection Between DSGE's and VAR's

- Fernandez-Villaverde, Rubio-Ramirez, Sargent, Watson Result
- Vector Autoregression

$$Y_t = B_1 Y_{t-1} + B_2 Y_{t-2} + ... + u_t,$$

  where u_t is iid.
- 'Matching impulse response functions' strategy for building DSGE models fits VARs and assumes u_t are a rotation of economic shocks (for details, see later notes).
- Can use the state space, observer representation to assess this assumption from the perspective of a DSGE.

Connection Between DSGE's and VAR's

- System (ignoring constant terms and measurement error):

$$\begin{aligned}(\text{'State equation'})\ \xi_t &= F\xi_{t-1} + D\epsilon_t,\ D = \begin{pmatrix} B \\ 0 \\ I \end{pmatrix}, \\ (\text{'Observer equation'})\ Y_t &= H\xi_t.\end{aligned}$$

- Substituting:

$$Y_t = HF\xi_{t-1} + HD\epsilon_t$$

- Suppose HD is square and invertible. Then

$$\epsilon_t = (HD)^{-1} Y_t - (HD)^{-1} HF\xi_{t-1}\ (**)$$

Substitute latter into the state equation:

$$\begin{aligned}\xi_t &= F\xi_{t-1} + D(HD)^{-1} Y_t - D(HD)^{-1} HF\xi_{t-1} \\ &= \left[I - D(HD)^{-1} H\right] F\xi_{t-1} + D(HD)^{-1} Y_t.\end{aligned}$$

Connection Between DSGE's and VAR's

We have:

$$\xi_t = M\xi_{t-1} + D\left(HD\right)^{-1} Y_t, \ M = \left[I - D\left(HD\right)^{-1} H\right] F.$$

If eigenvalues of M are less than unity,

$$\xi_t = D\left(HD\right)^{-1} Y_t + MD\left(HD\right)^{-1} Y_{t-1} + M^2 D\left(HD\right)^{-1} Y_{t-2} + ...$$

Substituting into $(**)$

$$\epsilon_t = \left(HD\right)^{-1} Y_t - \left(HD\right)^{-1} HF$$

$$\times \left[D\left(HD\right)^{-1} Y_{t-1} + MD\left(HD\right)^{-1} Y_{t-2} + M^2 D\left(HD\right)^{-1} Y_{t-3} + ...\right].$$

or,

$$Y_t = B_1 Y_{t-1} + B_2 Y_{t-2} + ... \ + u_t,$$

where

$$u_t = HD\epsilon_t, \ B_j = HFM^{j-1} D\left(HD\right)^{-1}, \ j = 1, 2, ...$$

- The latter is the VAR representation.

Connection Between DSGE's and VAR's

- The VAR repersentation is:

$$Y_t = B_1 Y_{t-1} + B_2 Y_{t-2} + \ldots + u_t,$$

where

$$u_t = HD\epsilon_t,\ B_j = HFM^{j-1}D\,(HD)^{-1},\ j = 1, 2, \ldots$$

- Notes:
  - ϵ_t is 'invertible' because it lies in space of current and past Y_t's.
  - VAR is *infinite*-ordered.
  - assumed system is 'square' (same number of elements in ϵ_t and Y_t). Sims-Zha (Macroeconomic Dynamics) show how to recover ϵ_t from current and past Y_t when the dimension of ϵ_t is greater than the dimension of Y_t.

Quick Review of Probability Theory

- Two random variables, $x \in (x_1, x_2)$ and $y \in (y_1, y_2)$.
- *Joint distribution*: $p(x, y)$

$$\begin{array}{c|c|c|} & x_1 & x_2 \\ \hline y_1 & p_{11} & p_{12} \\ \hline y_2 & p_{21} & p_{22} \\ \hline \end{array} = \begin{array}{c|c|c|} & x_1 & x_2 \\ \hline y_1 & 0.05 & 0.40 \\ \hline y_2 & 0.35 & 0.20 \\ \hline \end{array}$$

where

$$p_{ij} = probability\left(x = x_i, y = y_j\right).$$

- Restriction:

$$\int_{x,y} p(x, y)\, dxdy = 1.$$

Quick Review of Probability Theory

- *Joint distribution*: $p(x, y)$

$$\begin{array}{cc} & \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{array}{c} y_1 \\ y_2 \end{array} & \begin{array}{|c|c|} \hline p_{11} & p_{12} \\ \hline p_{21} & p_{22} \\ \hline \end{array} \end{array} = \begin{array}{cc} & \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{array}{c} y_1 \\ y_2 \end{array} & \begin{array}{|c|c|} \hline 0.05 & 0.40 \\ \hline 0.35 & 0.20 \\ \hline \end{array} \end{array}$$

- *Marginal distribution* of x : $p(x)$

Probabilities of various values of x without reference to the value of y:

$$p(x) = \begin{cases} p_{11} + p_{21} = 0.40 & x = x_1 \\ p_{12} + p_{22} = 0.60 & x = x_2 \end{cases}.$$

or,

$$p(x) = \int_y p(x, y)\, dy$$

Quick Review of Probability Theory

- *Joint distribution*: $p(x, y)$

$$\begin{array}{c|c|c|} & x_1 & x_2 \\ \hline y_1 & p_{11} & p_{12} \\ \hline y_2 & p_{21} & p_{22} \\ \hline \end{array} = \begin{array}{c|c|c|} & x_1 & x_2 \\ \hline y_1 & 0.05 & 0.40 \\ \hline y_2 & 0.35 & 0.20 \\ \hline \end{array}$$

- *Conditional distribution* of x given y : $p(x|y)$
  - Probability of x given that the value of y is known

$$p(x|y_1) = \begin{cases} p(x_1|y_1) & \frac{p_{11}}{p_{11}+p_{12}} = \frac{p_{11}}{p(y_1)} = \frac{0.05}{0.45} = 0.11 \\ p(x_2|y_1) & \frac{p_{12}}{p_{11}+p_{12}} = \frac{p_{12}}{p(y_1)} = \frac{0.40}{0.45} = 0.89 \end{cases}$$

or,

$$p(x|y) = \frac{p(x, y)}{p(y)}.$$

Quick Review of Probability Theory

- *Joint distribution*: $p(x, y)$

	x_1	x_2	
y_1	0.05	0.40	$p(y_1) = 0.45$
y_2	0.35	0.20	$p(y_2) = 0.55$
	$p(x_1) = 0.40$	$p(x_2) = 0.60$	

- Mode
  - *Mode of joint distribution* (in the example):

$$\text{argmax}_{x,y} p(x, y) = (x_2, y_1)$$

  - *Mode of the marginal distribution*:

$$\text{argmax}_x p(x) = x_2, \ \text{argmax}_y p(y) = y_2$$

  - Note: mode of the marginal and of joint distribution conceptually different.

Maximum Likelihood Estimation

- State space-observer system:

$$
\begin{aligned}
\xi_{t+1} &= F\xi_t + u_{t+1},\ Eu_t u_t' = Q, \\
Y_t^{data} &= a_0 + H\xi_t + w_t,\ Ew_t w_t' = R
\end{aligned}
$$

- Reduced form parameters, (F, Q, a_0, H, R), functions of θ.
- Choose θ to maximize likelihood, $p\left(Y^{data}|\theta\right)$:

$$
\begin{aligned}
p\left(Y^{data}|\theta\right) &= p\left(Y_1^{data}, \ldots, Y_T^{data}|\theta\right) \\
&= p\left(Y_1^{data}|\theta\right) \times p\left(Y_2^{data}|Y_1^{data}, \theta\right) \\
&\quad \times \cdots \times \overbrace{p\left(Y_t^{data}|Y_{t-1}^{data} \cdots Y_1^{data}, \theta\right)}^{\text{computed using Kalman Filter}} \\
&\quad \times \cdots \times p\left(Y_T^{data}|Y_{T-1}^{data}, \cdots, Y_1^{data}, \theta\right)
\end{aligned}
$$

- Kalman filter straightforward (see, e.g., Hamilton's textbook).

Bayesian Inference

- Bayesian inference is about describing the mapping from prior beliefs about θ, summarized in $p(\theta)$, to new posterior beliefs in the light of observing the data, Y^{data}.
- General property of probabilities:

$$p\left(Y^{data},\theta\right) = \begin{cases} p\left(Y^{data}|\theta\right) \times p\left(\theta\right) \\ p\left(\theta|Y^{data}\right) \times p\left(Y^{data}\right) \end{cases},$$

  which implies Bayes' rule:

$$p\left(\theta|Y^{data}\right) = \frac{p\left(Y^{data}|\theta\right) p\left(\theta\right)}{p\left(Y^{data}\right)},$$

  mapping from prior to posterior induced by Y^{data}.

Bayesian Inference

- Report features of the posterior distribution, $p\left(\theta | Y^{data}\right)$.
  - The value of θ that maximizes $p\left(\theta | Y^{data}\right)$, 'mode' of posterior distribution.
  - Compare marginal prior, $p\left(\theta_i\right)$, with marginal posterior of individual elements of θ, $g\left(\theta_i | Y^{data}\right)$:

$$g\left(\theta_i | Y^{data}\right) = \int_{\theta_{j \neq i}} p\left(\theta | Y^{data}\right) d\theta_{j \neq i} \text{ (multiple integration!!)}$$

  - Probability intervals about the mode of θ ('Bayesian confidence intervals'), need $g\left(\theta_i | Y^{data}\right)$.
- Marginal likelihood for assessing model 'fit':

$$p\left(Y^{data}\right) = \int_{\theta} p\left(Y^{data} | \theta\right) p\left(\theta\right) d\theta \text{ (multiple integration)}$$

Monte Carlo Integration: Simple Example

- Much of Bayesian inference is about multiple integration.
- Numerical methods for multiple integration:
  - Quadrature integration (example: approximating the integral as the sum of the areas of triangles beneath the integrand).
  - Monte Carlo Integration: uses random number generator.
- Example of Monte Carlo Integration:
  - suppose you want to evaluate

$$\int_a^b f(x)\,dx,\ -\infty \leq a < b \leq \infty.$$

  - select a density function, $g(x)$ for $x \in [a, b]$ and note:

$$\int_a^b f(x)\,dx = \int_a^b \frac{f(x)}{g(x)} g(x)\,dx = E\frac{f(x)}{g(x)},$$

  where E is the expectation operator, given $g(x)$.

Monte Carlo Integration: Simple Example

- Previous result: can express an integral as an expectation relative to a (arbitrary, subject to obvious regularity conditions) density function.
- Use the law of large numbers (LLN) to approximate the expectation.
  - step 1: draw x_i independently from density, g, for $i = 1, ..., M$.
  - step 2: evaluate $f(x_i) / g(x_i)$ and compute:

$$\mu_M \equiv \frac{1}{M} \sum_{i=1}^{M} \frac{f(x_i)}{g(x_i)} \rightarrow_{M \rightarrow \infty} E \frac{f(x)}{g(x)}.$$

- Exercise.
  - Consider an integral where you have an analytic solution available, e.g., $\int_0^1 x^2 dx$.
  - Evaluate the accuracy of the Monte Carlo method using various distributions on $[0, 1]$ like uniform or Beta.

Monte Carlo Integration: Simple Example

- Standard classical sampling theory applies.
- Independence of $f(x_i)/g(x_i)$ over i implies:

$$var\left(\frac{1}{M}\sum_{i=1}^{M}\frac{f(x_i)}{g(x_i)}\right) = \frac{v_M}{M},$$

$$v_M \equiv var\left(\frac{f(x_i)}{g(x_i)}\right) \simeq \frac{1}{M}\sum_{i=1}^{M}\left[\frac{f(x_i)}{g(x_i)} - \mu_M\right]^2.$$

- Central Limit Theorem
  - Estimate of $\int_a^b f(x)\,dx$ is a realization from a Nomal distribution with mean estimated by μ_M and variance, v_M/M.
  - With 95% probability,

$$\mu_M - 1.96 \times \sqrt{\frac{v_M}{M}} \;\leq\; \int_a^b f(x)\,dx \;\leq\; \mu_M + 1.96 \times \sqrt{\frac{v_M}{M}}$$

  - Pick g to minimize variance in $f(x_i)/g(x_i)$ and M to minimize (subject to computing cost) v_M/M.

Markov Chain, Monte Carlo (MCMC) Algorithms

- Among the top 10 algorithms "with the greatest influence on the development and practice of science and engineering in the 20th century".
  - Reference: January/February 2000 issue of Computing in Science & Engineering, a joint publication of the American Institute of Physics and the IEEE Computer Society.'

- Developed in 1946 by John von Neumann, Stan Ulam, and Nick Metropolis (see http://www.siam.org/pdf/news/637.pdf)

MCMC Algorithm: Overview

- compute a sequence, $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(M)}$, of values of the $N \times 1$ vector of model parameters in such a way that

$$\lim_{M \to \infty} Frequency\left[\theta^{(i)} \text{ close to } \theta\right] = p\left(\theta | Y^{data}\right).$$

- Use $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(M)}$ to obtain an approximation for
  - $E\theta$, $Var(\theta)$ under posterior distribution, $p\left(\theta | Y^{data}\right)$
  - $g\left(\theta^{i} | Y^{data}\right) = \int_{\theta_{i \neq j}} p\left(\theta | Y^{data}\right) d\theta d\theta$
  - $p\left(Y^{data}\right) = \int_{\theta} p\left(Y^{data} | \theta\right) p(\theta) d\theta$
  - posterior distribution of any function of θ, $f(\theta)$ (e.g., impulse responses functions, second moments).
- MCMC also useful for computing posterior mode, $\arg\max_{\theta} p\left(\theta | Y^{data}\right)$.

MCMC Algorithm: setting up

- Let $G(\theta)$ denote the log of the posterior distribution (excluding an additive constant):

$$G(\theta) = \log p\left(Y^{data}|\theta\right) + \log p(\theta);$$

- Compute posterior mode:

$$\theta^* = \arg\max_{\theta} G(\theta).$$

- Compute the positive definite matrix, V :

$$V \equiv \left[-\frac{\partial^2 G(\theta)}{\partial\theta\partial\theta'}\right]^{-1}_{\theta=\theta^*}$$

- Later, we will see that V is a rough estimate of the variance-covariance matrix of θ under the posterior distribution.

MCMC Algorithm: Metropolis-Hastings

- $\theta^{(1)} = \theta^*$
- to compute $\theta^{(r)}$, for $r > 1$
  - step 1: select candidate $\theta^{(r)}$, x,

$$\text{draw } \underbrace{x}_{N\times 1} \text{ from } \theta^{(r-1)} + \overbrace{k \times N\left(\underbrace{0}_{N\times 1}, V\right)}^{\text{'jump' distribution'}}, \; k \text{ is a scalar}$$

  - step 2: compute scalar, λ :

$$\lambda = \frac{p\left(Y^{data}|x\right) p\left(x\right)}{p\left(Y^{data}|\theta^{(r-1)}\right) p\left(\theta^{(r-1)}\right)}$$

  - step 3: compute $\theta^{(r)}$:

$$\theta^{(r)} = \begin{cases} \theta^{(r-1)} & \text{if } u > \lambda \\ x & \text{if } u < \lambda \end{cases}, \; u \text{ is a realization from uniform } [0,1]$$

Practical issues

- What is a sensible value for k?
  - set k so that you accept (i.e., $\theta^{(r)} = x$) in step 3 of MCMC algorithm are roughly 23 percent of time
- What value of M should you set?
  - want 'convergence', in the sense that if M is increased further, the econometric results do not change substantially
  - in practice, $M = 10,000$ (a small value) up to $M = 1,000,000$.
  - large M is time-consuming.
    - could use Laplace approximation (after checking its accuracy) in initial phases of research project.
    - more on Laplace below.
- Burn-in: in practice, some initial $\theta^{(i)}$'s are discarded to minimize the impact of initial conditions on the results.
- Multiple chains: may promote efficiency.
  - increase independence among $\theta^{(i)}$'s.
  - can do MCMC utilizing parallel computing (Dynare can do this).

MCMC Algorithm: Why Does it Work?

- Proposition that MCMC works may be surprising.
  - Whether or not it works does *not* depend on the details, i.e., precisely how you choose the jump distribution (of course, you had better use $k > 0$ and V positive definite).
    - Proof: see, e.g., Robert, C. P. (2001), *The Bayesian Choice*, Second Edition, New York: Springer-Verlag.
  - The details may matter by improving the efficiency of the MCMC algorithm, i.e., by influencing what value of M you need.
- Some Intuition
  - the sequence, $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(M)}$, is relatively heavily populated by θ's that have high probability and relatively lightly populated by low probability θ's.
  - Additional intuition can be obtained by positing a simple scalar distribution and using MATLAB to verify that MCMC approximates it well (see, e.g., question 2 in assignment 9).

MCMC Algorithm: using the Results

- To approximate marginal posterior distribution, $g\left(\theta_i | Y^{data}\right)$, of θ_i,
  - compute and display the histogram of $\theta_i^{(1)}, \theta_i^{(2)}, ..., \theta_i^{(M)}$, $i = 1, ..., M$.
- Other objects of interest:
  - mean and variance of posterior distribution θ :

$$E\theta \simeq \bar{\theta} \equiv \frac{1}{M} \sum_{j=1}^{M} \theta^{(j)},\ Var\left(\theta\right) \simeq \frac{1}{M} \sum_{j=1}^{M} \left[\theta^{(j)} - \bar{\theta}\right] \left[\theta^{(j)} - \bar{\theta}\right]'.$$

MCMC Algorithm: using the Results

- More complicated objects of interest:
  - impulse response functions,
  - model second moments,
  - forecasts,
  - Kalman smoothed estimates of real rate, natural rate, etc.
- All these things can be represented as non-linear functions of the model parameters, i.e., $f(\theta)$.
  - can approximate the distribution of $f(\theta)$ using

$$f\left(\theta^{(1)}\right), \ldots, f\left(\theta^{(M)}\right)$$

$$\rightarrow \quad Ef(\theta) \simeq \bar{f} \equiv \frac{1}{M}\sum_{i=1}^{M} f\left(\theta^{(i)}\right),$$

$$Var(f(\theta)) \quad \simeq \quad \frac{1}{M}\sum_{i=1}^{M}\left[f\left(\theta^{(i)}\right)-\bar{f}\right]\left[f\left(\theta^{(i)}\right)-\bar{f}\right]'$$

MCMC: Remaining Issues

- In addition to the first and second moments already discused, would also like to have the marginal likelihood of the data.
- Marginal likelihood is a Bayesian measure of model fit.

MCMC Algorithm: the Marginal Likelihood

- Consider the following sample average:

$$\frac{1}{M}\sum_{j=1}^{M}\frac{h\left(\theta^{(j)}\right)}{p\left(Y^{data}|\theta^{(j)}\right)p\left(\theta^{(j)}\right)},$$

  where $h(\theta)$ is an arbitrary density function over the $N-$ dimensional variable, θ.

By the law of large numbers,

$$\frac{1}{M}\sum_{j=1}^{M}\frac{h\left(\theta^{(j)}\right)}{p\left(Y^{data}|\theta^{(j)}\right)p\left(\theta^{(j)}\right)} \underset{M\to\infty}{\to} E\left(\frac{h(\theta)}{p\left(Y^{data}|\theta\right)p(\theta)}\right)$$

MCMC Algorithm: the Marginal Likelihood

$$\frac{1}{M}\sum_{j=1}^{M}\frac{h\left(\theta^{(j)}\right)}{p\left(Y^{data}|\theta^{(j)}\right)p\left(\theta^{(j)}\right)} \rightarrow_{M\rightarrow\infty} E\left(\frac{h\left(\theta\right)}{p\left(Y^{data}|\theta\right)p\left(\theta\right)}\right)$$

$$= \int_{\theta}\left(\frac{h\left(\theta\right)}{p\left(Y^{data}|\theta\right)p\left(\theta\right)}\right)\frac{p\left(Y^{data}|\theta\right)p\left(\theta\right)}{p\left(Y^{data}\right)}d\theta \;\; = \frac{1}{p\left(Y^{data}\right)}.$$

- When $h\left(\theta\right)=p\left(\theta\right)$, *harmonic mean estimator of the marginal likelihood*.
- Ideally, want an h such that the variance of

$$\frac{h\left(\theta^{(j)}\right)}{p\left(Y^{data}|\theta^{(j)}\right)p\left(\theta^{(j)}\right)}$$

  is small (recall the earlier discussion of Monte Carlo integration). More on this below.

Laplace Approximation to Posterior Distribution

- In practice, MCMC algorithm very time intensive.

- Laplace approximation is easy to compute and in many cases it provides a 'quick and dirty' approximation that is quite good.

Let $\theta \in R^N$ denote the $N-$dimensional vector of parameters and, as before,

$$G(\theta) \equiv \log p\left(Y^{data}|\theta\right) p(\theta)$$

$$p\left(Y^{data}|\theta\right) \text{ ~likelihood of data}$$

$$p(\theta) \text{ ~prior on parameters}$$

$$\theta^* \text{ ~maximum of } G(\theta) \text{ (i.e., mode)}$$

Laplace Approximation

Second order Taylor series expansion of
$G(\theta) \equiv \log\left[p\left(Y^{data}|\theta\right)p(\theta)\right]$ about $\theta = \theta^*$:

$$G(\theta) \approx G(\theta^*) + G_\theta(\theta^*)(\theta-\theta^*) - \frac{1}{2}(\theta-\theta^*)' G_{\theta\theta}(\theta^*)(\theta-\theta^*),$$

where

$$G_{\theta\theta}(\theta^*) = -\frac{\partial^2 \log p\left(Y^{data}|\theta\right)p(\theta)}{\partial\theta\partial\theta'}|_{\theta=\theta^*}$$

Interior optimality of θ^* implies:

$$G_\theta(\theta^*) = 0,\ G_{\theta\theta}(\theta^*) \text{ positive definite}$$

Then:

$$\begin{aligned} & p\left(Y^{data}|\theta\right)p(\theta) \\ \simeq\ & p\left(Y^{data}|\theta^*\right)p(\theta^*)\exp\left\{-\frac{1}{2}(\theta-\theta^*)' G_{\theta\theta}(\theta^*)(\theta-\theta^*)\right\}. \end{aligned}$$

Laplace Approximation to Posterior Distribution

Property of Normal distribution:

$$\int_\theta \frac{1}{(2\pi)^{\frac{N}{2}}} \left| G_{\theta\theta}(\theta^*) \right|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\theta - \theta^*)' G_{\theta\theta}(\theta^*) (\theta - \theta^*) \right\} d\theta = 1$$

Then,

$$\begin{aligned}
\int p\left(Y^{data}|\theta\right) p(\theta)\, d\theta &\simeq \int p\left(Y^{data}|\theta^*\right) p(\theta^*) \\
&\quad \times \exp\left\{ -\frac{1}{2} (\theta - \theta^*)' G_{\theta\theta}(\theta^*) (\theta - \theta^*) \right\} \\
&= \frac{p\left(Y^{data}|\theta^*\right) p(\theta^*)}{\frac{1}{(2\pi)^{\frac{N}{2}}} \left| G_{\theta\theta}(\theta^*) \right|^{\frac{1}{2}}}.
\end{aligned}$$

Laplace Approximation

- Conclude:

$$p\left(Y^{data}\right) \simeq \frac{p\left(Y^{data}|\theta^*\right)p\left(\theta^*\right)}{\frac{1}{(2\pi)^{\frac{N}{2}}}\left|G_{\theta\theta}\left(\theta^*\right)\right|^{\frac{1}{2}}}.$$

- Laplace approximation to posterior distribution:

$$\begin{aligned}\frac{p\left(Y^{data}|\theta\right)p\left(\theta\right)}{p\left(Y^{data}\right)} \quad &\simeq \quad \frac{1}{(2\pi)^{\frac{N}{2}}}\left|G_{\theta\theta}\left(\theta^*\right)\right|^{\frac{1}{2}} \\ &\quad \times \exp\left\{-\frac{1}{2}\left(\theta-\theta^*\right)' G_{\theta\theta}\left(\theta^*\right)\left(\theta-\theta^*\right)\right\}\end{aligned}$$

- So, posterior of θ_i (i.e., $g\left(\theta_i|Y^{data}\right)$) is approximately

$$\theta_i \sim N\left(\theta_i^*, \left[G_{\theta\theta}\left(\theta^*\right)^{-1}\right]_{ii}\right).$$

Modified Harmonic Mean Estimator of Marginal Likelihood

- Harmonic mean estimator of the marginal likelihood, $p\left(Y^{data}\right)$:

$$\left[\frac{1}{M}\sum_{j=1}^{M}\frac{h\left(\theta^{(j)}\right)}{p\left(Y^{data}|\theta^{(j)}\right)p\left(\theta^{(j)}\right)}\right]^{-1},$$

with $h\left(\theta\right)$ set to $p\left(\theta\right)$.

  - In this case, the marginal likelihood is the harmonic mean of the likelihood, evaluated at the values of θ generated by the MCMC algorithm.
  - Problem: the variance of the object being averaged is likely to be high, requiring high M for accuracy.

- When $h\left(\theta\right)$ is instead equated to Laplace approximation of posterior distribution, then $h\left(\theta\right)$ is approximately proportional to $p\left(Y^{data}|\theta^{(j)}\right)p\left(\theta^{(j)}\right)$ so that the variance of the variable being averaged in the last expression is low.

The Marginal Likelihood and Model Comparison

- Suppose we have two models, $Model$ 1 and $Model$ 2.
  - compute $p\left(Y^{data}|Model\ 1\right)$ and $p\left(Y^{data}|Model\ 2\right)$
- Suppose $p\left(Y^{data}|Model\ 1\right) > p\left(Y^{data}|Model\ 2\right)$. Then, posterior odds on Model 1 higher than Model 2.
  - 'Model 1 fits better than Model 2'
- Can use this to compare across two different models, or to evaluate contribution to fit of various model features: habit persistence, adjustment costs, etc.
  - For an application of this and the other methods in these notes, see Smets and Wouters, AER 2007.