Bayesian Inference for DSGE Models

Lawrence J. Christiano
• State space-observer form.
  – convenient for model estimation and many other things.

• Bayesian inference
  – Bayes’ rule.
  – Monte Carlo integration.
  – 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 \( x_t \):
    \[
    \hat{x}_t \equiv \frac{x_t - x}{x},
    \]
    \[
    \implies \frac{x_t}{x} = 1 + \hat{x}_t
    \]
    \[
    \implies \log \left( \frac{x_t}{x} \right) = \log (1 + \hat{x}_t) \approx \hat{x}_t
    \]

• Suppose we have a model solution in hand:\(^1\)
  \[
  z_t = Az_{t-1} + Bs_t
  \]
  \[
  s_t = Ps_{t-1} + \epsilon_t, \quad E\epsilon_t\epsilon_t' = D.
  \]

\(^1\)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, $N_t$:

$$
\begin{align*}
  z_t &= \left( \begin{array}{c} \hat{K}_{t+1} \\ \hat{N}_t \end{array} \right),
  s_t = \hat{\epsilon}_t, 
  \epsilon_t = e_t.
\end{align*}
$$

• 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:

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

so that

$$
  \hat{y}_t = \hat{\epsilon}_t + \alpha \hat{K}_t + (1 - \alpha) \hat{N}_t
  = \left( \begin{array}{cc} 0 & 1 - \alpha \\ \alpha & 0 \end{array} \right) z_t + \left( \begin{array}{c} 0 \\ 1 \end{array} \right) z_{t-1} + s_t
$$

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

$$
  \gamma_t^{data} = \left( \begin{array}{c} \log y_t \\ \log \hat{N}_t \end{array} \right) = \left( \begin{array}{c} \log y \\ \log \hat{N} \end{array} \right) + \left( \begin{array}{c} \hat{y}_t \\ \hat{N}_t \end{array} \right)
$$
State Space/Observer Form

• Model prediction for data:

\[
\gamma_{t}^{data} = \left( \frac{\text{log } y}{\text{log } N} \right) + \left( \hat{N}_{t} \right)
\]

\[
= \left( \frac{\text{log } y}{\text{log } N} \right) + \left[ \begin{array}{cc}
0 & 1 - \alpha \\
0 & 1 \\
\end{array} \right] z_{t} + \left[ \begin{array}{cc}
\alpha & 0 \\
0 & 0 \\
\end{array} \right] z_{t-1} + \left[ \begin{array}{c}
1 \\
0 \\
\end{array} \right] s_{t}
\]

\[
= a + H\zeta_{t}
\]

\[
\zeta_{t} = \begin{pmatrix}
z_{t} \\
z_{t-1} \\
\hat{\epsilon}_{t}
\end{pmatrix}, \quad a = \begin{pmatrix}
\text{log } y \\
\text{log } N
\end{pmatrix}, \quad H = \begin{pmatrix}
0 & 1 - \alpha & \alpha & 0 & 0 & 1 \\
0 & 1 & \alpha & 0 & 0 & 0
\end{pmatrix}
\]

• The *Observer Equation* may include measurement error, \( w_{t} \):

\[
\gamma_{t}^{data} = a + H\zeta_{t} + w_{t}, \quad Ew_{t}w_{t}^{\prime} = R.
\]

• Semantics: \( \zeta_{t} \) is the *state* of the system (do not confuse with the economic state \((K_{t}, \epsilon_{t})\)!).
State Space/Observer Form

- Law of motion of the state, $\xi_t$ (state-space equation):

$$
\begin{align*}
\dot{\xi}_t &= F\xi_{t-1} + u_t, \\
Eu_t u'_t &= Q
\end{align*}
$$

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

\[
u_t = \begin{pmatrix}
  B \\
  0 \\
  I
\end{pmatrix} \epsilon_t, \\
Q = \begin{pmatrix}
  BDB' & 0 & BD \\
  0 & 0 & 0 \\
  DB' & 0 & D
\end{pmatrix}, \\
F = \begin{pmatrix}
  A & 0 & BP \\
  I & 0 & 0 \\
  0 & 0 & P
\end{pmatrix}.
\]
Uses of State Space/Observer Form

- Estimation of $\theta$ and forecasting $\xi_t$ and $\gamma_{\text{data}}^t$
- 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):
    \[ P \left[ \xi_t | \gamma_{t-1}^{\text{data}}, \gamma_{t-2}^{\text{data}}, \ldots, \gamma_1^{\text{data}} \right], t = 1, 2, \ldots, T. \]
  - Smoothing:
    \[ P \left[ \xi_t | \gamma_T^{\text{data}}, \ldots, \gamma_1^{\text{data}} \right], t = 1, 2, \ldots, T. \]
  - Example: ‘real rate of interest’ and ‘output gap’ can be recovered from $\xi_t$ 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, \( t = 0, 1/3, 2/3, 1, 4/3, 5/3, 2, ... \)
  - suppose gdp is quarterly, \( t = 0, 1, 2, 3, .... \)

\[
\mathbf{Y}_t^{data} = \begin{pmatrix}
GDP_t \\
\text{monthly inflation}_t \\
\text{monthly inflation}_{t-1/3} \\
\text{hours}_t \\
\text{hours}_{t-1/3} \\
\text{hours}_{t-2/3}
\end{pmatrix}, \quad t = 0, 1, 2, ... .
\]

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

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

• Solution: easy!
Mixed Monthly/Quarterly Observations

- Model timing: $t = 0, 1/3, 2/3, ...$
  
  \[ z_t = A z_{t-1/3} + B s_t, \]
  \[ s_t = P s_{t-1/3} + e_t, \]
  \[ E e_t e_t' = D. \]

- Monthly state-space observer system, $t = 0, 1/3, 2/3, ...$
  
  \[ \bar{z}_t = F \bar{z}_{t-1/3} + u_t, \]
  \[ E u_t u_t' = Q, \]
  \[ u_t \sim iid \quad t = 0, 1/3, 2/3, ... \]

- Note:
  
  first order vector autoregressive representation for quarterly state

  \[ \bar{z}_t = F^3 \bar{z}_{t-1} + u_t + F u_{t-1/3} + F^2 u_{t-2/3}, \]

  \[ u_t + F u_{t-1/3} + F^2 u_{t-2/3} \sim iid \quad \text{for } t = 0, 1, 2, ...!\]
Mixed Monthly/Quarterly Observations

Consider the following system:

\[
\begin{pmatrix}
\zeta_t \\
\zeta_{t-\frac{1}{3}} \\
\zeta_{t-\frac{2}{3}}
\end{pmatrix} =
\begin{bmatrix}
F^3 & 0 & 0 \\
F^2 & 0 & 0 \\
F & 0 & 0
\end{bmatrix}
\begin{pmatrix}
\zeta_{t-1} \\
\zeta_{t-\frac{4}{3}} \\
\zeta_{t-\frac{5}{3}}
\end{pmatrix} +
\begin{bmatrix}
I & F & F^2 \\
0 & I & F \\
0 & 0 & I
\end{bmatrix}
\begin{pmatrix}
\tilde{u}_t \\
\tilde{u}_{t-\frac{1}{3}} \\
\tilde{u}_{t-\frac{2}{3}}
\end{pmatrix}.
\]

Define

\[
\tilde{\zeta}_t = \begin{pmatrix}
\zeta_t \\
\zeta_{t-\frac{1}{3}} \\
\zeta_{t-\frac{2}{3}}
\end{pmatrix}, \tilde{F} = \begin{bmatrix}
F^3 & 0 & 0 \\
F^2 & 0 & 0 \\
F & 0 & 0
\end{bmatrix}, \tilde{\tilde{u}}_t = \begin{bmatrix}
I & F & F^2 \\
0 & I & F \\
0 & 0 & I
\end{bmatrix}
\begin{pmatrix}
\tilde{u}_t \\
\tilde{u}_{t-\frac{1}{3}} \\
\tilde{u}_{t-\frac{2}{3}}
\end{pmatrix},
\]

so that

\[
\tilde{\zeta}_t = \tilde{F}\tilde{\zeta}_{t-1} + \tilde{\tilde{u}}_t, \quad \tilde{\tilde{u}}_t \sim iid \text{ in quarterly data, } t = 0, 1, 2, ...
\]

\[
E\tilde{\tilde{u}}_t\tilde{\tilde{u}}_t' = \tilde{\tilde{Q}} = \begin{bmatrix}
I & F & F^2 \\
0 & I & F \\
0 & 0 & I
\end{bmatrix} \begin{bmatrix}
D & 0 & 0 \\
0 & D & 0 \\
0 & 0 & D
\end{bmatrix} \begin{bmatrix}
I & F & F^2 \\
0 & I & F \\
0 & 0 & I
\end{bmatrix}'.
\]
Mixed Monthly/Quarterly Observations

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

\[
\tilde{\zeta}_t = \tilde{F}\tilde{\zeta}_{t-1} + \tilde{u}_t, \quad \tilde{u}_t \sim iid, \quad E\tilde{u}_t\tilde{u}'_t = \tilde{Q},
\]

\[
\gamma^\text{data}_t = \tilde{H}\tilde{\zeta}_t + w_t, \quad w_t \sim iid, \quad Ew_tw'_t = R.
\]

• Here, \( \tilde{H} \) selects elements of \( \tilde{\zeta}_t \) needed to construct \( \gamma^\text{data}_t \)
  
  – 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):

\[
(\text{‘State equation’}) \quad \xi_t = F\xi_{t-1} + D\epsilon_t, \quad D = \begin{pmatrix} B \\ 0 \\ I \end{pmatrix},
\]

\[
(\text{‘Observer equation’}) \quad Y_t = H\xi_t.
\]

- Substituting:

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

- Suppose \(HD\) is square and invertible. Then

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

Substitute latter into the state equation:

\[
\xi_t = F\xi_{t-1} + D (HD)^{-1}Y_t - D (HD)^{-1}HF\xi_{t-1}
\]

\[
= \left[ I - D (HD)^{-1}H \right] F\xi_{t-1} + D (HD)^{-1}Y_t.
\]
Connection Between DSGE’s and VAR’s

We have:

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

If eigenvalues of \( M \) are less than unity,

\[ \xi_t = D (HD)^{-1} Y_t + MD (HD)^{-1} Y_{t-1} + M^2 D (HD)^{-1} Y_{t-2} + \ldots \]

Substituting into (**)

\[ \epsilon_t = (HD)^{-1} Y_t - (HD)^{-1} HF \]
\[ \times \left[ D (HD)^{-1} Y_{t-1} + MD (HD)^{-1} Y_{t-2} + M^2 D (HD)^{-1} Y_{t-3} + \ldots \right] . \]

or,

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

where

\[ u_t = H D \epsilon_t, \quad B_j = H F M^{j-1} D (HD)^{-1}, \quad j = 1, 2, \ldots \]

• The latter is the VAR representation.
Connection Between DSGE’s and VAR’s

- The VAR representation is:

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

where

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

- Notes:
  - \( \epsilon_t \) is ‘invertible’ because it lies in space of current and past \( Y_t \)’s.
  - VAR is \textit{infinite}-ordered.
  - assumed system is ‘square’ (same number of elements in \( \epsilon_t \) and \( Y_t \)). Sims-Zha (Macroeconomic Dynamics) show how to recover \( \epsilon_t \) from current and past \( Y_t \) when the dimension of \( \epsilon_t \) is greater than the dimension of \( Y_t \).
• Two random variables, $x \in (x_1, x_2)$ and $y \in (y_1, y_2)$.

• Joint distribution: $p(x, y)$

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

where

\[
p_{ij} = \text{probability } (x = x_i, y = y_j).
\]

• Restriction:

\[
\int_{x,y} p(x, y) \, dxdy = 1.
\]
Quick Review of Probability Theory

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

  
<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_1$</td>
<td>$p_{11}$</td>
<td>$p_{12}$</td>
</tr>
<tr>
<td>$y_2$</td>
<td>$p_{21}$</td>
<td>$p_{22}$</td>
</tr>
</tbody>
</table>

  $y_1 = p_{11} + p_{21} = 0.05 + 0.35 = 0.40$
  $y_2 = p_{12} + p_{22} = 0.20$

- **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**: \( p(x, y) \)

\[
egin{array}{c|cc}
  & x_1 & x_2 \\
\hline
y_1 & p_{11} & p_{12} \\
 y_2 & p_{21} & p_{22}
\end{array}
\]

\[
egin{array}{c|cc}
 x_1 & x_2 \\
\hline
 y_1 & 0.05 & 0.40 \\
 y_2 & 0.35 & 0.20
\end{array}
\]

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

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

or,

\[
p(x|y) = \frac{p(x, y)}{p(y)}.
\]
Quick Review of Probability Theory

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

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$p(y_1)$</th>
<th>$p(y_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_1$</td>
<td>0.05</td>
<td>0.40</td>
<td>0.45</td>
<td>0.55</td>
</tr>
<tr>
<td>$y_2$</td>
<td>0.35</td>
<td>0.20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Mode**
  - Mode of joint distribution (in the example):
    \[
    \arg\max_{x, y} p(x, y) = (x_2, y_1)
    \]
  - Mode of the marginal distribution:
    \[
    \arg\max_x p(x) = x_2, \quad \arg\max_y p(y) = y_2
    \]
  - Note: mode of the marginal and of joint distribution conceptually different.
Maximum Likelihood Estimation

• State space-observer system:

\[
\begin{align*}
\zeta_{t+1} &= F\zeta_t + u_{t+1}, \quad Eu_t u'_t = Q, \\
Y_{t}^{data} &= a_0 + H\zeta_t + w_t, \quad Ew_t w'_t = R
\end{align*}
\]

• Reduced form parameters, \((F, Q, a_0, H, R)\), functions of \(\theta\).

• Choose \(\theta\) to maximize likelihood, \(p\left(Y_{\text{data}} | \theta \right)\):

\[
p\left(Y_{\text{data}} | \theta \right) = p\left(Y_{\text{data}}^1, ..., Y_{\text{data}}^T | \theta \right) \\
= p\left(Y_{\text{data}}^1 | \theta \right) \times p\left(Y_{\text{data}}^2 | Y_{\text{data}}^1, \theta \right) \\
\times \cdots \times p\left(Y_{\text{data}}^T | Y_{\text{data}}^{T-1} \cdots Y_{\text{data}}^1, \theta \right) \\
\times \cdots \times p\left(Y_{\text{data}}^T | Y_{\text{data}}^{T-1}, \cdots, Y_{\text{data}}^1, \theta \right)
\]

• Kalman filter straightforward (see, e.g., Hamilton’s textbook).
Bayesian Inference

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

• General property of probabilities:

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

which implies Bayes’ rule:

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

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

• Report features of the posterior distribution, \( p(\theta|Y^{data}) \).
  
  – The value of \( \theta \) that maximizes \( p(\theta|Y^{data}) \), ‘mode’ of posterior distribution.
  
  – Compare marginal prior, \( p(\theta_i) \), with marginal posterior of individual elements of \( \theta \), \( g(\theta_i|Y^{data}) \):

\[
g(\theta_i|Y^{data}) = \int_{\theta_j \neq i} p(\theta|Y^{data}) \, d\theta_{j \neq i} \text{ (multiple integration!!)}
\]

  – Probability intervals about the mode of \( \theta \) (‘Bayesian confidence intervals’), need \( g(\theta_i|Y^{data}) \).

• Marginal likelihood for assessing model ‘fit’:

\[
p(Y^{data}) = \int_{\theta} p(Y^{data}|\theta) \, p(\theta) \, d\theta \text{ (multiple integration)}
\]
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
  \[ \int_{a}^{b} f(x) \, dx, \quad -\infty \leq a < b \leq \infty. \]
  - 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 \( g(x) \).

> Integral from a to b of f(x) dx, minus infinity <= a < b <= infinity.
> Select a density function, g(x) for x in [a, b] and note:
> Integral from a to b of f(x) dx = Integral from a to b of f(x)/g(x) * g(x) dx = E(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:

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

- 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:

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

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

- Central Limit Theorem
  - Estimate of $\int_{a}^{b} f(x) \, dx$ is a realization from a Normal distribution with mean estimated by $\mu_M$ and variance, $\nu_M/M$.
  - With 95% probability,

$$\mu_M - 1.96 \times \sqrt{\frac{\nu_M}{M}} \leq \int_{a}^{b} f(x) \, dx \leq \mu_M + 1.96 \times \sqrt{\frac{\nu_M}{M}}$$

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

• 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, \( \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)} \), of values of the \( N \times 1 \) vector of model parameters in such a way that

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

• Use \( \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)} \) to obtain an approximation for
  - \( E \theta, \text{Var} \left( \theta \right) \) under posterior distribution, \( p \left( \theta | Y^{\text{data}} \right) \)
  - \( g \left( \theta_i | Y^{\text{data}} \right) = \int_{\theta_i \neq j} p \left( \theta | Y^{\text{data}} \right) \, d\theta d\theta \)
  - \( p \left( Y^{\text{data}} \right) = \int_{\theta} p \left( Y^{\text{data}} | \theta \right) p \left( \theta \right) \, d\theta \)
  - posterior distribution of any function of \( \theta, f \left( \theta \right) \) (e.g., impulse responses functions, second moments).

• MCMC also useful for computing posterior mode, \( \arg \max_{\theta} p \left( \theta | Y^{\text{data}} \right) \).
MCMC Algorithm: setting up

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

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

- 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 $\theta$ under the posterior distribution.
MCMC Algorithm: Metropolis-Hastings

• \(\theta^{(1)} = \theta^*\)
• to compute \(\theta^{(r)}\), for \(r > 1\)
  – step 1: select candidate \(\theta^{(r)}, x,\)
    
    \[
    \text{draw } x \text{ from } \theta^{(r-1)} + k \times N \begin{pmatrix} 0 \\ N \times 1 \end{pmatrix}, \text{ } k \text{ is a scalar}
    \]

  – step 2: compute scalar, \(\lambda:\)
    
    \[
    \lambda = \frac{p \left( Y_{\text{data}} | x \right) p \left( x \right)}{p \left( Y_{\text{data}} | \theta^{(r-1)} \right) p \left( \theta^{(r-1)} \right)}
    \]

  – step 3: compute \(\theta^{(r)}:\)
    
    \[
    \theta^{(r)} = \begin{cases} 
    \theta^{(r-1)} & \text{if } u > \lambda \\
    x & \text{if } u < \lambda
    \end{cases}, \text{ } u \text{ is a realization from uniform } [0, 1]
    \]
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).
  – 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, \( \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)} \), is relatively heavily populated by \( \theta \)'s that have high probability and relatively lightly populated by low probability \( \theta \)'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 \( \theta_i \),
  - 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 \( \theta \) :
    \[
    E\theta \sim \bar{\theta} \equiv \frac{1}{M} \sum_{j=1}^{M} \theta^{(j)}, \quad Var(\theta) \sim \frac{1}{M} \sum_{j=1}^{M} \left[ \theta^{(j)} - \bar{\theta} \right] \left[ \theta^{(j)} - \bar{\theta} \right]',
    \]
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

\[
\begin{align*}
 f(\theta^{(1)}),&..., f(\theta^{(M)}) \\
 \rightarrow \quad Ef(\theta) &\approx \bar{f} \equiv \frac{1}{M} \sum_{i=1}^{M} f(\theta^{(i)}), \\
 Var(f(\theta)) &\approx \frac{1}{M} \sum_{i=1}^{M} \left[ f(\theta^{(i)}) - \bar{f} \right] \left[ f(\theta^{(i)}) - \bar{f} \right]^\prime
\end{align*}
\]
MCMC: Remaining Issues

- In addition to the first and second moments already discussed, would also like to have the marginal likelihood of the data.
- Marginal likelihood is a Bayesian measure of model fit.
Consider the following sample average:

\[
\frac{1}{M} \sum_{j=1}^{M} \frac{h(\theta^{(j)})}{p(Y_{data}|\theta^{(j)}) p(\theta^{(j)})}
\]

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

By the law of large numbers,

\[
\frac{1}{M} \sum_{j=1}^{M} \frac{h(\theta^{(j)})}{p(Y_{data}|\theta^{(j)}) p(\theta^{(j)})} \rightarrow_{M \to \infty} E \left( \frac{h(\theta)}{p(Y_{data}|\theta) p(\theta)} \right)
\]
MCMC Algorithm: the Marginal Likelihood

\[ \frac{1}{M} \sum_{j=1}^{M} \frac{h(\theta^{(j)})}{p(Y_{\text{data}}|\theta^{(j)})p(\theta^{(j)})} \xrightarrow{M \to \infty} E \left( \frac{h(\theta)}{p(Y_{\text{data}}|\theta)p(\theta)} \right) \]

\[ = \int_{\theta} \left( \frac{h(\theta)}{p(Y_{\text{data}}|\theta)p(\theta)} \right) \frac{p(Y_{\text{data}}|\theta)p(\theta)}{p(Y_{\text{data}})} d\theta = \frac{1}{p(Y_{\text{data}})}. \]

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

\[ \frac{h(\theta^{(j)})}{p(Y_{\text{data}}|\theta^{(j)})p(\theta^{(j)})} \]

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 $\theta \in \mathbb{R}^N$ denote the $N$–dimensional vector of parameters and, as before,

\[
G(\theta) \equiv \log p \left( \gamma^{data} | \theta \right) p(\theta)
\]

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

$p(\theta)$ \~ prior on parameters

$\theta^*$ \~ maximum of $G(\theta)$ (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

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

Interior optimality of \( \theta^* \) implies:

\[ G_\theta(\theta^*) = 0, \quad G_{\theta \theta}(\theta^*) \] positive definite

Then:

\[ p \left( Y^{data} | \theta \right) p(\theta) \]

\[ \simeq p \left( Y^{data} | \theta^* \right) p(\theta^*) \exp \left\{ -\frac{1}{2} (\theta - \theta^*)' G_{\theta \theta}(\theta^*) (\theta - \theta^*) \right\} . \]
Laplace Approximation to Posterior Distribution

Property of Normal distribution:

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

Then,

$$\int p \left( Y_{data} | \theta \right) p \left( \theta \right) d\theta \simeq \int p \left( Y_{data} | \theta^* \right) p \left( \theta^* \right) \times \exp \left\{ -\frac{1}{2} (\theta - \theta^*)' G_{\theta\theta} (\theta^*) (\theta - \theta^*) \right\}$$

$$= \frac{p \left( Y_{data} | \theta^* \right) p \left( \theta^* \right)}{\frac{1}{(2\pi)^{N/2}} |G_{\theta\theta} (\theta^*)|^{1/2}}.$$
Laplace Approximation

- Conclude:

\[
p(Y_{\text{data}}) \approx \frac{p(Y_{\text{data}}|\theta^*) p(\theta^*)}{\frac{1}{(2\pi)^{N/2}} |G_{\theta\theta}(\theta^*)|^{1/2}}.
\]

- Laplace approximation to posterior distribution:

\[
\frac{p(Y_{\text{data}}|\theta) p(\theta)}{p(Y_{\text{data}})} \approx \frac{1}{(2\pi)^{N/2}} |G_{\theta\theta}(\theta^*)|^{1/2} \times \exp \left\{ -\frac{1}{2} (\theta - \theta^*)' G_{\theta\theta}(\theta^*) (\theta - \theta^*) \right\}
\]

- So, posterior of \( \theta_i \) (i.e., \( g(\theta_i|Y_{\text{data}}) \)) is approximately

\[
\theta_i \sim N \left( \theta_i^*, \left[ G_{\theta\theta}(\theta^*)^{-1} \right]_{ii} \right).
\]
Modified Harmonic Mean Estimator of Marginal Likelihood

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

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

with \( h (\theta) \) set to \( p (\theta) \).
- In this case, the marginal likelihood is the harmonic mean of the likelihood, evaluated at the values of \( \theta \) generated by the MCMC algorithm.
- Problem: the variance of the object being averaged is likely to be high, requiring high \( M \) for accuracy.

- When \( h (\theta) \) is instead equated to Laplace approximation of posterior distribution, then \( h (\theta) \) 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 \( p(\mathcal{Y} | \text{Model 1}) \) and \( p(\mathcal{Y} | \text{Model 2}) \)

- Suppose \( p(\mathcal{Y} | \text{Model 1}) > p(\mathcal{Y} | \text{Model 2}) \). 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.