Outline

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

• Preliminaries.
  – Probabilities.
  – Maximum Likelihood.

• Bayesian inference
  – Bayes’ rule.
  – Bayesians versus Classical.
  – Monte Carlo integration.
  – MCMC algorithm.
  – Laplace approximation.
  – Marginal Likelihood of the Data.
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 \( x_t \):
    \[
    \hat{x}_t \equiv \frac{x_t - x}{x},
    \]
    \[
    \Rightarrow \frac{x_t}{x} = 1 + \hat{x}_t
    \]
    \[
    \Rightarrow \log\left(\frac{x_t}{x}\right) = \log(1 + \hat{x}_t) \approx \hat{x}_t
    \]

- 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, $N_t$:

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

- 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:

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

so that

$$\hat{y}_t = \hat{\epsilon}_t + \alpha \hat{K}_t + (1 - \alpha) \hat{N}_t$$

$$= \begin{pmatrix} 0 & 1 - \alpha \end{pmatrix} z_t + \begin{pmatrix} \alpha & 0 \end{pmatrix} z_{t-1} + s_t$$

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

$$\gamma_{t}^{data} = \begin{pmatrix} \log y_t \\ \log \hat{N}_t \end{pmatrix} = \begin{pmatrix} \log y \\ \log \hat{N} \end{pmatrix} + \begin{pmatrix} \hat{y}_t \\ \hat{N}_t \end{pmatrix}$$
State Space/Observer Form

- Model prediction for data:

\[
\gamma_{\text{data}}^t = \left( \begin{array}{c} \log y \\ \log N \end{array} \right) + \left( \begin{array}{c} \hat{y}_t \\ \hat{N}_t \end{array} \right) \\
= \left( \begin{array}{c} \log y \\ \log N \end{array} \right) + \begin{bmatrix} 0 & 1 - \alpha \\ 0 & 1 \end{bmatrix} z_t + \begin{bmatrix} \alpha & 0 \\ 0 & 0 \end{bmatrix} z_{t-1} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} s_t \\
= a + H\xi_t
\]

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

- The Observer Equation may include measurement error, \( w_t \):

\[
\gamma_{\text{data}}^t = a + H\xi_t + w_t, \quad Ew_t w'_t = R.
\]

- Semantics: \( \xi_t \) 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, $\bar{\xi}_t$ (state-space equation):

$$\bar{\xi}_t = F\bar{\xi}_{t-1} + u_t, \quad Eu_t u'_t = Q$$

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

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

\[ \xi_t = F \xi_{t-1} + u_t, \quad Eu_t u'_t = Q, \]

\[ Y^\text{data}_t = a + H \xi_t + \omega_t, \quad E\omega_t \omega'_t = R. \]

- Can be constructed from model parameters

\[ \theta = (\beta, \delta, \ldots) \]

so

\[ F = F(\theta), \quad Q = Q(\theta), \quad a = a(\theta), \quad H = H(\theta), \quad R = R(\theta). \]
Uses of State Space/Observer Form

- Estimation of $\theta$ and forecasting $\xi_t$ 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):
    $$P \left[ \xi_t | Y_{t-1}^{data}, Y_{t-2}^{data}, \ldots, Y_1^{data} \right], \ t = 1, 2, \ldots, T.$$  
  - Smoothing:
    $$P \left[ \xi_t | Y_T^{data}, \ldots, Y_1^{data} \right], \ t = 1, 2, \ldots, T.$$  
  - 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
Quick Review of Probability Theory

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

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

where

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

- Restriction:

$$\int_{x,y} p(x,y) \, dx \, dy = 1.$$
Quick Review of Probability Theory

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

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

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

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

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

  \[
  \begin{array}{ccc}
  & x_1 & x_2 \\
  y_1 & 0.05 & 0.40 & p(y_1) = 0.45 \\
  y_2 & 0.35 & 0.20 & p(y_2) = 0.55 \\
  \end{array}
  \]

  $p(x_1) = 0.40$ $p(x_2) = 0.60$

- **Mode**
  - *Mode of joint distribution* (in the example):
    \[
    \arg\max_{x, y} p(x, y) = (x_2, y_1)
    \]
  
  - *Mode of the marginal distribution*:
    \[
    \arg\max_x p(x) = x_2, \; \arg\max_y p(y) = y_2
    \]

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

- State space-observer system:

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

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

- Choose \(\theta\) to maximize likelihood, \(p\left(\gamma^\text{data}_j | \theta \right)\):

\[
p\left(\gamma^\text{data}_j | \theta \right) = p\left(\gamma^\text{data}_1, ..., \gamma^\text{data}_T | \theta \right) = p\left(\gamma^\text{data}_1 | \theta \right) \times p\left(\gamma^\text{data}_2 | \gamma^\text{data}_1, \theta \right) \times \cdots \times p\left(\gamma^\text{data}_T | \gamma^\text{data}_{T-1} \cdots \gamma^\text{data}_1, \theta \right)
\]

computed using Kalman Filter

- 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:

$$p(\theta|Y_{data}) = \frac{p(Y_{data}|\theta)p(\theta)}{p(Y_{data})},$$

mapping from prior to posterior induced by $Y_{data}$. 

Bayesians versus Classicals

- Maximum likelihood estimator and measure of model fit:

\[
\hat{\theta}_T = \arg \max_{\theta} p(Y|\theta) \\
\hat{L}_T = \log p(Y|\hat{\theta}_T)
\]

- Bayesian ‘estimate’ (‘maximum likelihood estimation with a penalty function’) and measure of model fit:

\[
\theta^* = \arg \max_{\theta} p(Y|\theta) p(\theta) \\
\log p(Y) = \log \left[ \int_\theta p(Y|\theta) p(\theta) \, d\theta \right]
\]
Bayesians versus Classicals: Example

- Suppose $\theta \in \{\theta_1, \theta_2\}$ and
  
  $p(Y|\theta_1) = 1$, $p(Y|\theta_2) = 0.2$,
  
  $p(\theta_1) = 0.1$, $p(\theta_2) = 0.9$.

- Classicals’ estimate and fit:
  
  ‘estimate’: $\hat{\theta}_T = \theta_1$, ‘fit’: $\hat{\mathcal{L}}_T = \log 1 = 0$

- Bayesians’ estimate and fit:

  $\theta^* = \theta_2 = \arg\max_{\{\theta_1, \theta_2\}} \left\{ \frac{0.1}{p(Y|\theta_1) p(\theta_1)}, \frac{0.18}{p(Y|\theta_2) p(\theta_2)} \right\}$

  $\log p(Y) = \log \left[ 1 \times 0.1 + 0.2 \times 0.9 \right] = -1.27 \ll \hat{\mathcal{L}}_T$. 
Bayesians versus Classicals

- Numerical example:
  
  \[ p(Y|\theta_1) = 1, \ p(Y|\theta_2) = 0.2, \]
  \[ p(\theta_1) = 0.1, \ p(\theta_2) = 0.9. \]

  - Classical chooses \( \theta = \theta_1 \), to get best possible fit.
  - For Bayesian, fit implied by \( \theta = \theta_1 \) of little interest because it requires a completely implausible value of \( \theta \).

- Example (see Christiano-Eichenbaum-Trabandt ECTA2016).
  
  - Diamond-Mortensen-Pissarides (DMP) model.
    
    - \( \theta_1 (\sim 0.3), \theta_2 (\sim 0.95) \) low and high values for wage replacement ratio, \( \theta \), respectively.

  - Classical: Hagedorn-Manovskii (AER2008), conclude DMP model is good because it fits aggregate data well with \( \theta = \theta_2 \).
  - Bayesian: Shimer (AER2005), concludes DMP model bad on grounds that \( \theta = \theta_2 \) is highly implausible.
Bayesian Inference

- Report features of the posterior distribution, $p(\theta | Y^{\text{data}})$.
  - The value of $\theta$ that maximizes $p(\theta | Y^{\text{data}})$, ‘mode’ of posterior distribution.
  - Compare marginal prior, $p(\theta_i)$, with marginal posterior of individual elements of $\theta$, $g(\theta_i | Y^{\text{data}})$:
    \[
g(\theta_i | Y^{\text{data}}) = \int_{\theta_{j\neq i}} p(\theta | Y^{\text{data}}) \, d\theta_{j\neq i} \quad \text{(multiple integration!!)}
    \]
  - Probability intervals about the mode of $\theta$ (‘Bayesian confidence intervals’), need $g(\theta_i | Y^{\text{data}})$.

- Marginal likelihood for assessing model ‘fit’:
  \[
p(Y^{\text{data}}) = \int_{\theta} p(Y^{\text{data}} | \theta) \, p(\theta) \, d\theta \quad \text{(multiple integration)}
  \]
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 \( g(x) \).
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, $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(M)}$, of values of the $N \times 1$ vector of model parameters in such a way that

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

- Use $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(M)}$ to obtain an approximation for
  - $E \theta$, $Var(\theta)$ under posterior distribution, $p\left( \theta \mid Y^{data} \right)$
  - $g\left( \theta^i \mid Y^{data} \right) = \int_{\theta \neq i} p\left( \theta \mid Y^{data} \right) d\theta d\theta$
  - $p\left( Y^{data} \right) = \int_{\theta} p\left( Y^{data} \mid \theta \right) p\left( \theta \right) d\theta$
  - 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, $\arg \max_{\theta} p\left( \theta \mid Y^{data} \right)$. 
MCMC Algorithm: setting up

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

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

- Compute posterior mode:

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

- Compute the positive definite matrix, $V$:

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

- 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 $\theta^{(r)}$, $x$,
    
    \[
    \text{draw } x \text{ from } \theta^{(r-1)} + k \times N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, V \right), k \text{ is a scalar}
    \]
    
  - step 2: compute scalar, $\lambda$
    
    \[
    \lambda = \frac{p \left( Y^{data} | x \right) p \left( x \right)}{p \left( Y^{data} | \theta^{(r-1)} \right) p \left( \theta^{(r-1)} \right)}
    \]
  
  - step 3: compute $\theta^{(r)}$
    
    \[
    \theta^{(r)} = \begin{cases} 
    \theta^{(r-1)} & \text{if } u > \lambda \\
    x & \text{if } u < \lambda 
    \end{cases}, \text{ } u \text{ is a realization from uniform } [0, 1]
    \]
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, $\theta^{(1)}, \theta^{(2)},..., \theta^{(M)}$, 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 in assignment related to this lecture).
Why a Low Acceptance Rate is Desirable
MCMC Algorithm: using the Results

- To approximate marginal posterior distribution, \( g \left( \theta_i | Y_{\text{data}} \right) \), of \( \theta_i \),
  - compute and display the histogram of \( \theta_i^{(1)}, \theta_i^{(2)}, \ldots, \theta_i^{(M)} \),
    \( i = 1, \ldots, M \).

- Other objects of interest:
  - mean and variance of posterior distribution \( \theta \):
    \[
    E\theta \simeq \bar{\theta} \equiv \frac{1}{M} \sum_{j=1}^{M} \theta^{(j)}, \quad \text{Var} (\theta) \simeq \frac{1}{M} \sum_{j=1}^{M} \left[ \theta^{(j)} - \bar{\theta} \right] \left[ \theta^{(j)} - \bar{\theta} \right]' .
    \]
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

\[
\begin{align*}
  f(\theta^{(1)}), & \ldots, f(\theta^{(M)}) \\
  \rightarrow \quad Ef(\theta) & \approx \bar{f} \equiv \frac{1}{M} \sum_{i=1}^{M} f(\theta^{(i)}), \\
  \text{Var}(f(\theta)) & \approx \frac{1}{M} \sum_{i=1}^{M} [f(\theta^{(i)}) - \bar{f}] [f(\theta^{(i)}) - \bar{f}]',
\end{align*}
\]
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.
MCMC Algorithm: the Marginal Likelihood

- Consider the following sample average:

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

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

By the law of large numbers,

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

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

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

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

$$G (\theta) \equiv \log p \left( Y_{\text{data}} | \theta \right) p (\theta)$$

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

$p (\theta)$ ~prior on parameters

$\theta^* \sim$maximum of $G (\theta)$ (i.e., mode)
Laplace Approximation

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

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

Then:

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

Property of Normal distribution:

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

Then,

\[
\int p\left(Y^{data} | \theta \right) p\left( \theta \right) d\theta \approx \int p\left(Y^{data} | \theta^* \right) p\left( \theta^* \right) \times \exp \left\{ -\frac{1}{2} (\theta - \theta^*)' G_{\theta\theta}(\theta^*) (\theta - \theta^*) \right\} d\theta
\]

\[
= \frac{p\left(Y^{data} | \theta^* \right) p\left( \theta^* \right)}{1 \frac{1}{(2\pi)^\frac{N}{2}} |G_{\theta\theta}(\theta^*)|^\frac{1}{2}}.
\]
Laplace Approximation

- Conclude:
  \[
  p \left( Y^{data} \right) \approx \frac{p \left( Y^{data} | \theta^* \right) p \left( \theta^* \right)}{\frac{1}{(2\pi)^{N/2}} |G_{\theta\theta} (\theta^*)|^{1/2}}.
  \]

- Laplace approximation to posterior distribution:
  \[
  \frac{p \left( Y^{data} | \theta \right) p \left( \theta \right)}{p \left( Y^{data} \right)} \approx \frac{1}{(2\pi)^{N/2}} |G_{\theta\theta} (\theta^*)|^{1/2}
  \times \exp \left\{ -\frac{1}{2} (\theta - \theta^*)' G_{\theta\theta} (\theta^*) (\theta - \theta^*) \right\}
  \]

- So, posterior of \( \theta_i \) (i.e., \( g \left( \theta_i | Y^{data} \right) \)) is approximately
  \[
  \theta_i \sim N \left( \theta_i^*, \left[ G_{\theta\theta} (\theta^*)^{-1} \right]_{ii} \right).
  \]
Figure 16 Priors and posteriors of estimated parameters of the medium-sized DSGE model.
Modified Harmonic Mean Estimator of Marginal Likelihood

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

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

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 \( p(Y^{data}|\theta^{(j)}) p(\theta^{(j)}) \) 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 $p(Y^\text{data}|\text{Model 1})$ and $p(Y^\text{data}|\text{Model 2})$

• Suppose $p(Y^\text{data}|\text{Model 1}) > p(Y^\text{data}|\text{Model 2})$. 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.