Bayesian Inference for DSGE Models

Lawrence J. Christiano

Outline

- State space-observer form.
  - convenient for model estimation and many other things.
- Preliminaries.
  - Probabilities.
  - Maximum Likelihood.
- Bayesian inference
  - Bayes' rule.
  - Bayesians versus Classicals.
  - Monte Carlo integation.
  - 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 :

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

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

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

[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):

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

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

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

State Space/Observer Form

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

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

- Can be constructed from model parameters

$$\theta = (\beta, \delta, ...)$$

so

$$F = F(\theta),\ Q = Q(\theta),\ a = a(\theta),\ H = H(\theta),\ R = R(\theta).$$

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

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*: $p(x, y)$

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

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

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

- *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):

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

  - *Mode of the marginal distribution*:

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

  - 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, $p\left(Y^{data}|\theta\right)$:

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

- 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:

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

which implies Bayes' rule:

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

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

Bayesians versus Classicals

- Maximum likelihood estimator and measure of model fit:

$$\begin{aligned} \hat{\theta}_T &= \arg\max_{\theta} p\left(Y|\theta\right) \\ \hat{\mathcal{L}}_T &= \log p\left(Y|\hat{\theta}_T\right) \end{aligned}$$

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

$$\begin{aligned} \theta^* &= \arg\max_{\theta} p\left(Y|\theta\right) p\left(\theta\right) \\ \log p\left(Y\right) &= \log\left[\int_{\theta} p\left(Y|\theta\right) p\left(\theta\right) d\theta\right] \end{aligned}$$

Bayesians versus Classicals: Example

- Suppose $\theta \in \{\theta_1, \theta_2\}$ and

$$\begin{aligned} p(Y|\theta_1) &= 1,\ p(Y|\theta_2) = 0.2, \\ p(\theta_1) &= 0.1,\ p(\theta_2) = 0.9. \end{aligned}$$

- Classicals' estimate and fit:

$$\text{'estimate': } \hat{\theta}_T = \theta_1,\ \text{'fit': } \hat{\mathcal{L}}_T = \log 1 = 0$$

- Bayesians' estimate and fit:

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

$$\log p(Y) = \log[1 \times 0.1 + 0.2 \times 0.9] = -1.27 \lll \hat{\mathcal{L}}_T.$$

Bayesians versus Classicals

- Numerical example:

$$\begin{aligned} p(Y|\theta_1) &= 1,\ p(Y|\theta_2) = 0.2, \\ p(\theta_1) &= 0.1,\ p(\theta_2) = 0.9. \end{aligned}$$

  - 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 θ.

- Example (see Christiano-Eichenbaum-Trabandt ECTA2016).
  - Diamond-Mortensen-Pissarides (DMP) model.
    - $\theta_1 (\simeq 0.3)$, $\theta_2 (\simeq 0.95)$ low and high values for wage replacement ratio, θ, 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, $p\left(\theta|Y^{data}\right)$.
  - 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,

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

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

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 in assignment related to this lecture).

Why a Low Acceptance Rate is Desirable

MCMC Algorithm: using the Results

- To approximate marginal posterior distribution, $g\left(\theta_i | Y^{data}\right)$, 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

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

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

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

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

By the law of large numbers,

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

MCMC Algorithm: the Marginal Likelihood

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

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

- 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,

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

Laplace Approximation

- Conclude:

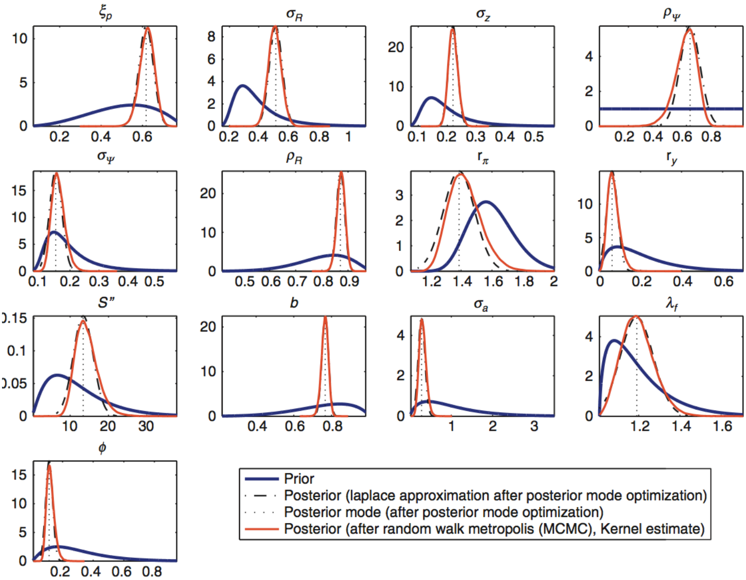
$$p\left(Y^{data}\right) \simeq \frac{p\left(Y^{data}|\theta^*\right)p\left(\theta^*\right)}{\frac{1}{(2\pi)^{\frac{N}{2}}}\left|G_{\theta\theta}\left(\theta^*\right)\right|^{\frac{1}{2}}}.$$

- 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

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



gure 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, $p\left(Y^{data}\right)$:

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

  with $h(\theta)$ set to $p(\theta)$.
  - 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 $h(\theta)$ 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 $p\left(Y^{data}|Model\ 1\right)$ 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.