Bayesian Maximum Likelihood

- Bayesians describe 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:

\[
p (Y_{data}, \theta) = \begin{cases} 
    p (Y_{data}|\theta) \times p (\theta) & , \\
    p (\theta|Y_{data}) \times p (Y_{data}) & ,
\end{cases}
\]

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 \( Y_{data} \).
• Properties of the posterior distribution, $p\left(\theta|Y^{data}\right)$.

  – The value of $\theta$ that maximizes $p\left(\theta|Y^{data}\right)$ (‘mode’ of posterior distribution).

  – Graphs that compare the marginal posterior distribution of individual elements of $\theta$ with the corresponding prior.

  – Probability intervals about the mode of $\theta$ (‘Bayesian confidence intervals’)

  – Other properties of $p\left(\theta|Y^{data}\right)$ helpful for assessing model ‘fit’. 
Bayesian Maximum Likelihood ...

- Computation of mode sometimes referred to as ‘Bayesian maximum likelihood’:

\[
\theta^{\text{mode}} = \arg \max_{\theta} \left\{ \log \left[ p \left( Y^{\text{data}} | \theta \right) \right] + \sum_{i=1}^{N} \log \left[ p_i \left( \theta_i \right) \right] \right\}
\]

maximum likelihood with a penalty function.

- Shape of posterior distribution, \( p \left( \theta | Y^{\text{data}} \right) \), obtained by Metropolis-Hastings algorithm.
  - Algorithm computes
    \[
    \theta (1), \ldots, \theta (N),
    \]
    which, as \( N \to \infty \), has a density that approximates \( p \left( \theta | Y^{\text{data}} \right) \) well.

  - Marginal posterior distribution of any element of \( \theta \) displayed as the histogram of the corresponding element \( \{ \theta (i), i = 1, \ldots, N \} \)
Metropolis-Hastings Algorithm (MCMC)

- We have (except for a constant):

\[
f \left( \begin{pmatrix} \theta \\ N \times 1 \end{pmatrix} | Y \right) = \frac{f (Y|\theta) f (\theta)}{f (Y)}.
\]

- We want the marginal posterior distribution of \( \theta_i \):

\[
h (\theta_i | Y) = \int_{\theta_j \neq i} f (\theta | Y) d\theta_j, \ i = 1, \ldots, N.
\]

- MCMC algorithm can approximate \( h (\theta_i | Y) \).

- Obtain (\( V \) produced automatically by gradient-based maximization methods):

\[
\theta^{\text{mode}} \equiv \theta^* = \arg \max_{\theta} f (Y|\theta) f (\theta), \ V \equiv \left[ -\frac{\partial^2 f (Y|\theta) f (\theta)}{\partial \theta \partial \theta'} \right]^{-1}_{\theta=\theta^*}.
\]
Metropolis-Hastings Algorithm (MCMC) ...

- Compute the sequence, \( \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)} \) (\( M \) large) whose distribution turns out to have pdf \( f(\theta|Y) \).

\(- \theta^{(1)} = \theta^* \)

- to compute \( \theta^{(r)} \), for \( r > 1 \)

  * step 1: select candidate \( \theta^{(r)}, x, \)

    draw \( x \) from \( \theta^{(r-1)} + kN \begin{pmatrix} \mathbf{0}_{N \times 1} \end{pmatrix}, V \), \( k \) is a scalar

  * step 2: compute scalar, \( \lambda \):

  \[ \lambda = \frac{f(Y|x) f(x)}{f(Y|\theta^{(r-1)}) f(\theta^{(r-1)})} \]

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

  \[ \theta^{(r)} = \begin{cases} 
  \theta^{(r-1)} & \text{if } u > \lambda \\
  x & \text{if } u < \lambda 
  \end{cases} \]

  \( u \) is a realization from uniform \([0, 1]\)
Metropolis-Hastings Algorithm (MCMC) ...

- Approximating marginal posterior distribution, $h(\theta_i | Y)$, of $\theta_i$
  
  - compute and display the histogram of $\theta^{(1)}_i, \theta^{(2)}_i, \ldots, \theta^{(M)}_i, i = 1, \ldots, N$.

- Other objects of interest:
  
  - mean and variance of posterior distribution $\theta$ :

    $$E\theta \sim \bar{\theta} \equiv \frac{1}{M} \sum_{j=1}^{M} \theta^{(j)}, \quad Var(\theta) \sim \frac{1}{M} \sum_{j=1}^{M} \left[ \theta^{(j)} - \bar{\theta} \right] \left[ \theta^{(j)} - \bar{\theta} \right]^\prime.$$
Metropolis-Hastings Algorithm (MCMC) ...

• Some intuition

  – Algorithm is more likely to select moves into high probability regions than into low probability regions.

  – Set, \( \{ \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)} \} \), populated relatively more by elements near mode of \( f(\theta|Y) \).

  – Set, \( \{ \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)} \} \), also populated (though less so) by elements far from mode of \( f(\theta|Y) \).
Metropolis-Hastings Algorithm (MCMC) ...

- Practical issues

  - what value should you set $k$ to?

    * set $k$ so that you accept (i.e., $\theta^{(r)} = x$) in step 3 of MCMC algorithm are roughly 27 percent of time

  - what value of $M$ should you set?

    * a value so that if $M$ is increased further, your results do not change

      · 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.
Laplace Approximation to Posterior Distribution

• In practice, Metropolis-Hasting algorithm very time intensive. Do it last!

• In practice, Laplace approximation is quick, essentially free and very accurate.

• Let $\theta \in \mathbb{R}^N$ denote the $N$–dimensional vector of parameters and

$$g(\theta) \equiv \log f(y|\theta) f(\theta),$$

$f(y|\theta) \sim$ likelihood of data

$f(\theta) \sim$ prior on parameters

$\theta^* \sim$ maximum of $g(\theta)$ (i.e., mode)
Laplace Approximation to Posterior Distribution ...

- Second order Taylor series expansion 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

$$g_{\theta\theta}(\theta^*) = -\frac{\partial^2 \log f(y|\theta) f(\theta)}{\partial \theta \partial \theta'}|_{\theta=\theta^*}$$

- Interior optimality implies:

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

- Then,

$$f(y|\theta) f(\theta) \approx f(y|\theta^*) f(\theta^*) \exp \left\{ -\frac{1}{2}(\theta - \theta^*)' g_{\theta\theta}(\theta^*)(\theta - \theta^*) \right\}.$$
Laplace Approximation to Posterior Distribution ...

• Note

\[
\frac{1}{(2\pi)^{N/2}} |g_{\theta\theta}(\theta^*)|^{1/2} \exp \left\{ -\frac{1}{2} (\theta - \theta^*)' g_{\theta\theta}(\theta^*) (\theta - \theta^*) \right\}
\]

= multinormal density for \( N \) - dimensional random variable \( \theta \)

with mean \( \theta^* \) and variance \( g_{\theta\theta}(\theta^*)^{-1} \).

• So, posterior of \( \theta_i \) (i.e., \( h(\theta_i|Y) \)) is approximately

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

• This formula for the posterior distribution is essentially free, because \( g_{\theta\theta} \) is computed as part of gradient-based numerical optimization procedures.
Laplace Approximation to Posterior Distribution ...

- Marginal likelihood of data, $y$, is useful for model comparisons. Easy to compute using the Laplace approximation.

- 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,

$$
\int f(y|\theta) f(\theta) d\theta \simeq \int f(y|^\theta) f(\theta^*) \exp \left\{ -\frac{1}{2} (\theta - \theta^*)' g_{\theta\theta}(\theta^*) (\theta - \theta^*) \right\} d\theta
$$

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

$$
= \frac{f(y|^\theta^*) f(\theta^*)}{\frac{1}{(2\pi)^{\frac{N}{2}}} |g_{\theta\theta}(\theta^*)|^\frac{1}{2}}.
$$
Laplace Approximation to Posterior Distribution ...

- Formula for marginal likelihood based on Laplace approximation:

\[
f(y) = \int f(y|\theta) f(\theta) d\theta \simeq (2\pi)^{-N/2} \frac{f(y|\theta^*) f(\theta^*)}{|g_{\theta\theta}(\theta^*)|^{1/2}}.
\]

- Suppose \( f(y|\text{Model 1}) > f(y|\text{Model 2}) \). 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.
Generalized Method of Moments

• Express your econometric estimator into Hansen’s GMM framework and you get standard errors

  – Essentially, \textit{any} estimation strategy fits (see Hamilton)

• Works when parameters of interest, \( \beta \), have the following property:

\[
E \begin{bmatrix} u_t \\ N \times 1 \end{bmatrix} \begin{bmatrix} \beta \\ n \times 1 \end{bmatrix} = 0, \ \beta \ \text{true value of some parameter(s) of interest}
\]

  \( u_t(\beta) \sim \text{stationary stochastic process} \) (and other conditions)

  – \( n = N \) : ‘exactly identified’

  – \( n < N \) : ‘over identified’
Generalized Method of Moments ... 

– Example 1: mean

\[ \beta = \mathbb{E}x_t, \]

\[ u_t(\beta) = \beta - x_t. \]

– Example 2: mean and variance

\[ \beta = \begin{bmatrix} \mu & \sigma \end{bmatrix}, \]

\[ \mathbb{E}x_t = \mu, \mathbb{E}(x_t - \mu)^2 = \sigma^2. \]

then,

\[ u_t(\beta) = \begin{bmatrix} \mu - x_t \\ (x_t - \mu)^2 - \sigma^2 \end{bmatrix}. \]
Generalized Method of Moments ...

– Example 3: mean, variance, correlation, relative standard deviation

\[ \beta = \begin{bmatrix} \mu_y & \sigma_y & \mu_x & \sigma_x & \rho_{xy} & \lambda \end{bmatrix}, \quad \lambda \equiv \sigma_x / \sigma_y, \]

where

\[ E y_t = \mu_y, \quad E (y_t - \mu_y)^2 = \sigma_y^2 \]

\[ E x_t = \mu_x, \quad E (x_t - \mu_x)^2 = \sigma_x^2 \]

\[ \rho_{xy} = \frac{E (y_t - \mu_y)(x_t - \mu_x)}{\sigma_y \sigma_x}. \]

then

\[ u_t (\beta) = \begin{bmatrix} \mu_x - x_t \\
(x_t - \mu_x)^2 - \sigma_x^2 \\
\mu_y - y_t \\
(y_t - \mu_y)^2 - \sigma_y^2 \\
\sigma_y \sigma_x \rho_{xy} - (y_t - \mu_y)(x_t - \mu_x) \\
\sigma_y \lambda - \sigma_x \end{bmatrix}. \]
Generalized Method of Moments ...

– Example 4: New Keynesian Phillips curve

\[ \pi_t = 0.99 E_t \pi_{t+1} + \gamma s_t, \]

or,

\[ \pi_t - 0.99 \pi_{t+1} - \gamma s_t = \eta_{t+1} \]

where,

\[ \eta_{t+1} = 0.99 (E_t \pi_{t+1} - \pi_{t+1}) \implies E_t \eta_{t+1} = 0 \]

Under Rational Expectations : \( \eta_{t+1} \perp \text{time } t \text{ information, } z_t \)

\[ u_t (\gamma) = [\pi_t - 0.99 \pi_{t+1} - \gamma s_t] z_t \]
Generalized Method of Moments ...

• Inference about $\beta$

  – Estimator of $\beta$ in exactly identified case ($n = N$)

    * Choose $\hat{\beta}$ to mimick population property of true $\beta$,

    $$E u_t (\beta) = 0.$$

    * Define:

    $$g_T (\beta) = \frac{1}{T} \sum_{t=1}^{T} u_t (\beta).$$

    * Solve

    $$\hat{\beta} : g_T \left( \begin{array}{c} \hat{\beta} \\ N \times 1 \end{array} \right) = 0, \begin{array}{c} N \times 1 \end{array}.$$
Generalized Method of Moments ... 

– Example 1: mean

\[ \beta = E x_t, \]

\[ u_t(\beta) = \beta - x_t. \]

Choose \( \hat{\beta} \) so that

\[ g_T(\hat{\beta}) = \frac{1}{T} \sum_{t=1}^{T} u_t(\hat{\beta}) = \hat{\beta} - \frac{1}{T} \sum_{t=1}^{T} x_t = 0 \]

and \( \hat{\beta} \) is simply sample mean.
Generalized Method of Moments ...

– Example 4 in exactly identified case

\[ E u_t (\gamma) = E \left[ \pi_t - 0.99\pi_{t+1} - \gamma s_t \right] z_t, \ z_t \sim \text{scalar} \]

choose \( \hat{\gamma} \) so that

\[ g_T \left( \hat{\beta} \right) = \frac{1}{T} \sum_{t=1}^{T} \left[ \pi_t - 0.99\pi_{t+1} - \hat{\gamma} s_t \right] z_t = 0, \]

or. standard instrumental variables estimator:

\[ \hat{\gamma} = \frac{\frac{1}{T} \sum_{t=1}^{T} \left[ \pi_t - 0.99\pi_{t+1} \right] z_t}{\frac{1}{T} \sum_{t=1}^{T} s_t z_t} \]
Generalized Method of Moments ...

– Key message:

☆ In exactly identified case, GMM does not deliver a new estimator you would not have thought of on your own

- means, correlations, regression coefficients, exactly identified IV estimation, maximum likelihood.

☆ GMM provides framework for deriving asymptotically valid formulas for estimating sampling uncertainty.
Generalized Method of Moments ...

- Estimating $\beta$ in overidentified case ($N > n$)
  * Cannot exactly implement sample analog of $E u_t (\beta) = 0$:

\[
g_T \begin{pmatrix} \hat{\beta} \\ n \times 1 \end{pmatrix} = \begin{pmatrix} 0 \\ N \times 1 \end{pmatrix}
\]

* Instead, ‘do the best you can’:

\[
\hat{\beta} = \arg \min_\beta g_T (\beta)' W_T g_T (\beta),
\]

where

$W_T \sim$ is a positive definite weighting matrix.

* GMM works for any positive definite $W_T$, but is most efficient if $W_T$ is inverse of estimator of variance-covariance matrix of $g_T (\hat{\beta})$:

\[
(W_T)^{-1} = E g_T (\hat{\beta}) g_T (\hat{\beta})'.
\]
Generalized Method of Moments ... 

– This choice of weighting matrix very sensible:

* weight heavily those moment conditions (i.e., elements of $g_T(\hat{\beta})$) that are precisely estimated

* pay less attention to the others.
Generalized Method of Moments ...

– Estimator of $W_{T}^{-1}$

* Note:

$$E g_T \left( \hat{\beta} \right) g_T \left( \hat{\beta} \right)'$$

$$= \frac{1}{T^2} E \left[ u_1 \left( \hat{\beta} \right) + u_2 \left( \hat{\beta} \right) + \ldots + u_T \left( \hat{\beta} \right) \right] \left[ u_1 \left( \hat{\beta} \right) + u_2 \left( \hat{\beta} \right) + \ldots + u_T \left( \hat{\beta} \right) \right]'$$

$$= \frac{1}{T} \left[ \frac{T}{T} E u_t \left( \hat{\beta} \right) u_t \left( \hat{\beta} \right)' + \frac{T-1}{T} E u_t \left( \hat{\beta} \right) u_{t+1} \left( \hat{\beta} \right)' + \ldots + \frac{1}{T} E u_t \left( \hat{\beta} \right) u_{t+T-1} \left( \hat{\beta} \right)' \right. $$

$$\left. + \frac{T-1}{T} E u_t \left( \hat{\beta} \right) u_{t-1} \left( \hat{\beta} \right)' + \frac{T-2}{T} E u_t \left( \hat{\beta} \right) u_{t-2} \left( \hat{\beta} \right)' + \ldots + \frac{1}{T} E u_t \left( \hat{\beta} \right) u_{t-T+1} \left( \hat{\beta} \right)' \right]$$

$$= \frac{1}{T} \left[ C (0) + \sum_{r=1}^{T-1} \frac{T-r}{T} \left( C (r) + C (r)' \right) \right],$$

where

$$C (r) = E u_t \left( \hat{\beta} \right) u_{t-r} \left( \hat{\beta} \right)'$$

$* W_{T}^{-1}$ is $\frac{1}{T} \times$ spectral density matrix at frequency zero, $S_0$, of $u_t \left( \hat{\beta} \right)'$.
Generalized Method of Moments ...

– Conclude:

\[ W_T^{-1} = Eg_T \left( \hat{\beta} \right) g_T \left( \hat{\beta} \right) = \frac{1}{T} \left[ C(0) + \sum_{r=1}^{T-1} \frac{T-r}{T} \left( C(r) + C(r)' \right) \right] = \frac{S_0}{T}. \]

– \( W_T^{-1} \) estimated by

\[ \hat{W}_T^{-1} = \frac{1}{T} \left[ \hat{C}(0) + \sum_{r=1}^{T-1} \frac{T-r}{T} \left( \hat{C}(r) + \hat{C}(r)' \right) \right] = \frac{1}{T} \hat{S}_0, \]

imposing whatever restrictions are implied by the null hypothesis, i.e., (as in ex. 4)

\[ C(r) = 0, \ r > R \text{ some } R. \]

– which is ‘Newey-West estimator of spectral density at frequency zero’

* Problem: need \( \hat{\beta} \) to compute \( W_T^{-1} \) and need \( W_T^{-1} \) to compute \( \hat{\beta}!! \)

· Solution - first compute \( \hat{\beta} \) using \( W_T = I \), then iterate...
Generalized Method of Moments ...

• Sampling Uncertainty in $\hat{\beta}$.

  – The exactly identified case

  – By the Mean Value Theorem, $g_T(\hat{\beta})$ can be expressed as follows:

$$g_T(\hat{\beta}) = g_T(\beta_0) + D(\hat{\beta} - \beta_0),$$

where $\beta_0$ is the true value of the parameters and

$$D = \left. \frac{\partial g_T(\beta)}{\partial \beta'} \right|_{\beta = \beta^*}, \text{ some } \beta^* \text{ between } \beta_0 \text{ and } \hat{\beta}.$$

  – Since $g_T(\hat{\beta}) = 0$ and $g_T(\beta_0) \sim N(0, S_0/T)$, it follows:

$$\hat{\beta} - \beta_0 = -D^{-1}g_T(\beta_0),$$

so

$$\hat{\beta} - \beta_0 \sim N \left(0, \frac{(D'S_0^{-1}D)^{-1}}{T} \right)$$
Generalized Method of Moments ...

– The overidentified case.

* An extension of the ideas we have already discussed.

* Can derive the results for yourself, using the ‘delta function method’ for deriving the sampling distribution of statistics.

* Hamilton’s text book has a great review of GMM.