Bayesian Maximum Likelihood

- Bayesians describe 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}.

Bayesian Maximum Likelihood ...

- Properties 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).
  - Graphs that compare the marginal posterior distribution of individual elements of θ with the corresponding prior.
  - Probability intervals about the mode of θ ('Bayesian confidence intervals')
  - Other properties of $p\left(\theta|Y^{data}\right)$ helpful for assessing model 'fit'.

- Computation of mode sometimes referred to as 'Basyesian maximum likelihood':

$$\theta^{\mathrm{mod}\,e} = \arg\max_{\theta} \left\{ \log\left[p\left(Y^{data} | \theta \right) \right] + \sum_{i=1}^{N} \log\left[p_i\left(\theta_i \right) \right] \right\}$$

  maximum likelihood with a penalty function.

- Shape of posterior distribution, $p\left(\theta | Y^{data}\right)$, obtained by Metropolis-Hastings algorithm.
  - Algorithm computes

$$\theta\left(1\right), ..., \theta\left(N\right),$$

    which, as $N \rightarrow \infty$, has a density that approximates $p\left(\theta | Y^{data}\right)$ well.

  - Marginal posterior distribution of any element of θ displayed as the histogram of the corresponding element $\left\{\theta\left(i\right), i = 1, .., N\right\}$

Metropolis-Hastings Algorithm (MCMC)

- We have (except for a constant):

$$f\left(\underbrace{\theta}_{N\times 1}|Y\right)=\frac{f\left(Y|\theta\right)f\left(\theta\right)}{f\left(Y\right)}.$$

- We want the marginal posterior distribution of θ_i :

$$h\left(\theta_i|Y\right)=\int_{\theta_{j\neq i}}f\left(\theta|Y\right)d\theta_{j\neq i},\ i=1,...,N.$$

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

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

$$\theta^{\mathrm{mod}\,e}\equiv\theta^{*}=\arg\max_{\theta}f\left(Y|\theta\right)f\left(\theta\right),\ V\equiv\left[-\frac{\partial^2 f\left(Y|\theta\right)f\left(\theta\right)}{\partial\theta\partial\theta'}\right]^{-1}_{\theta=\theta^*}.$$

- Compute the sequence, $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(M)}$ (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,

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

    * step 2: compute scalar, λ :

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

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

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

- Approximating marginal posterior distribution, $h\left(\theta_i|Y\right)$, of θ_i

  – compute and display the histogram of $\theta_i^{(1)}, \theta_i^{(2)}, ..., \theta_i^{(M)}, i = 1, ..., N.$

- Other objects of interest:

  – mean and variance of posterior distribution θ :

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

- Some intuition

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

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

  – Set, $\left\{\theta^{(1)}, \theta^{(2)}, ..., \theta^{(M)}\right\}$, also populated (though less so) by elements far from mode of $f\left(\theta|Y\right)$.

- 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 $\theta \in R^N$ denote the $N-$dimensional vector of parameters and

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

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

$f(\theta)$ ~prior on parameters

θ^* ~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,

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

- Note

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

$=$ multinormal density for $N-$ dimensional random variable θ

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

- So, posterior of θ_i (i.e., $h\left(\theta_i|Y\right)$) is approximately

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

- This formula for the posterior distribution is essentially free, because $g_{\theta\theta}$ is computed as part of gradient-based numerical optimization procedures.

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

- Property of Normal distribution:

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

- Then,

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

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

$$= \frac{f(y|\theta^*) f(\theta^*)}{\frac{1}{(2\pi)^{\frac{N}{2}}} \left| g_{\theta\theta}(\theta^*) \right|^{\frac{1}{2}}}.$$

- Formula for marginal likelihood based on Laplace approximation:

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

- Suppose $f(y|Model\ 1) > f(y|Model\ 2)$. 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, *any* estimation strategy fits (see Hamilton)
- Works when parameters of interest, β, have the following property:

$$E \underbrace{u_t}_{N\times 1}\left(\underbrace{\beta}_{n\times 1}\right) = 0,\ \beta \text{ true value of some parameter(s) of interest}$$

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

  - $n = N$: 'exactly identified'
  - $n < N$: 'over identified'

– Example 1: mean

$$\beta = Ex_t,$$

$$u_t(\beta) = \beta - x_t.$$

– Example 2: mean and variance

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

$$Ex_t = \mu, E(x_t - \mu)^2 = \sigma^2.$$

then,

$$u_t(\beta) = \begin{bmatrix} \mu - x_t \\ (x_t - \mu)^2 - \sigma^2 \end{bmatrix}.$$

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

$$\beta = \left[\begin{array}{cccccc} \mu_y & \sigma_y & \mu_x & \sigma_x & \rho_{xy} & \lambda \end{array}\right], \; \lambda \equiv \sigma_x/\sigma_y,$$

where

$$Ey_t = \mu_y, \; E\left(y_t - \mu_y\right)^2 = \sigma_y^2$$

$$Ex_t = \mu_x, \; E\left(x_t - \mu_x\right)^2 = \sigma_x^2$$

$$\rho_{xy} = \frac{E\left(y_t - \mu_y\right)\left(x_t - \mu_x\right)}{\sigma_y \sigma_x}.$$

then

$$u_t(\beta) = \left[\begin{array}{c} \mu_x - x_t \\ \left(x_t - \mu_x\right)^2 - \sigma_x^2 \\ \mu_y - y_t \\ \left(y_t - \mu_y\right)^2 - \sigma_y^2 \\ \sigma_y \sigma_x \rho_{xy} - \left(y_t - \mu_y\right)\left(x_t - \mu_x\right) \\ \sigma_y \lambda - \sigma_x \end{array}\right].$$

– 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,

$$\eta_{t+1} = 0.99\left(E_t\pi_{t+1} - \pi_{t+1}\right) \Longrightarrow E_t\eta_{t+1} = 0$$

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

$$u_t(\gamma) = \left[\pi_t - 0.99\pi_{t+1} - \gamma s_t\right] z_t$$

- Inference about β
  - Estimator of β in exactly identified case ($n = N$)
    - Choose $\hat{\beta}$ to mimick population property of true β,

$$Eu_t(\beta) = 0.$$

    - Define:

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

    - Solve

$$\hat{\beta} : \ g_T\left(\underbrace{\hat{\beta}}_{N\times 1}\right) = \underbrace{0}_{N\times 1}.$$

– Example 1: mean

$$\beta = Ex_t,$$

$$u_t(\beta) = \beta - x_t.$$

Choose $\hat{\beta}$ so that

$$g_T\left(\hat{\beta}\right) = \frac{1}{T}\sum_{t=1}^{T} u_t\left(\hat{\beta}\right) = \hat{\beta} - \frac{1}{T}\sum_{t=1}^{T} x_t = 0$$

and $\hat{\beta}$ is simply sample mean.

– Example 4 in exactly identified case

$$Eu_t(\gamma) = E[\pi_t - 0.99\pi_{t+1} - \gamma s_t] z_t, \; z_t \sim \text{scalar}$$

choose $\hat{\gamma}$ so that

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

or. standard instrumental variables estimator:

$$\hat{\gamma} = \frac{\frac{1}{T}\sum_{t=1}^{T}[\pi_t - 0.99\pi_{t+1}] z_t}{\frac{1}{T}\sum_{t=1}^{T} s_t z_t}$$

– 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.

– Estimating β in overidentified case ($N > n$)

* Cannot exactly implement sample analog of $Eu_t(\beta) = 0$:

$$g_T\left(\underbrace{\hat{\beta}}_{n\times 1}\right) = \underbrace{0}_{N\times 1}$$

* Instead, 'do the best you can':

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

where

W_T ~ 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 $g_T\left(\hat{\beta}\right)$:

$$(W_T)^{-1} = Eg_T\left(\hat{\beta}\right) g_T\left(\hat{\beta}\right)'.$$

– This choice of weighting matrix very sensible:

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

  - ∗ pay less attention to the others.

– Estimator of W_T^{-1}

$*$ Note:

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

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

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

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

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

where

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

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

– 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

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

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...

- Sampling Uncertainty in $\hat{\beta}$.
  - The exactly identified case
  - By the Mean Value Theorem, $g_T\left(\hat{\beta}\right)$ can be expressed as follows:

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

where β_0 is the true value of the parameters and

$$D = \frac{\partial g_T\left(\beta\right)}{\partial \beta'}|_{\beta=\beta^*},$$ some β^* between β_0 and $\hat{\beta}$.

  - Since $g_T\left(\hat{\beta}\right) = 0$ and $g_T\left(\beta_0\right) \overset{a}{\sim} N\left(0, S_0/T\right)$, it follows:

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

so

$$\hat{\beta} - \beta_0 \overset{a}{\sim} N\left(0, \frac{\left(D' S_0^{-1} D\right)^{-1}}{T}\right)$$

- 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.