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:

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

  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 conf dence intervals')
  - Other properties of $p\left(\theta|Y^{data}\right)$ helpful for assessing model 'f t'.

Bayesian Maximum Likelihood ...

- 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 $N \to \infty$, 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):

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

Metropolis-Hastings Algorithm (MCMC) ...

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

  – $\theta^{(1)} = \theta^{*}$

  – 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 :

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

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

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

Metropolis-Hastings Algorithm (MCMC) ...

- 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 :

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

  –

—

Metropolis-Hastings Algorithm (MCMC) ...

- 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 $f\left(\theta|Y\right)$.

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 $\theta \in R^N$ 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,

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

Laplace Approximation to Posterior Distribution ...

- 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.

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:

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

- Then,

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

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

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

Laplace Approximation to Posterior Distribution ...

- 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 f ts better than Model 2'

- Can use this to compare across two different models, or to evaluate contribution to f t of various model features: habit persistence, adjustment costs, etc.