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:

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

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

26
Bayesian Maximum Likelihood ...

- Properties of the posterior distribution, \( p(\theta|Y_{data}) \).
  - The value of \( \theta \) that maximizes \( p(\theta|Y_{data}) \) (‘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 \( p(\theta|Y_{data}) \) 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 \( \theta \left( 1 \right), \ldots, \theta \left( N \right) \),

  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 \left( i \right), i = 1, \ldots, N \}\}
Metropolis-Hastings Algorithm (MCMC)

- We have (except for a constant):
  \[
  f \left( \theta \in \mathbb{R}^N \mid Y \right) = \frac{f(Y \mid \theta) f(\theta)}{f(Y)}.
  \]

- We want the marginal posterior distribution of \( \theta_i \):
  \[
  h(\theta_i \mid Y) = \int_{\theta_j \neq i} f(\theta \mid Y) \, d\theta_j, \quad i = 1, \ldots, N.
  \]

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

- Obtain (\( V \) produced automatically by gradient-based maximization methods):
  \[
  \theta^{\text{mode}} \equiv \theta^* = \arg \max_{\theta} f(Y \mid \theta) f(\theta), \quad V \equiv \left[ -\frac{\partial^2 f(Y \mid \theta) f(\theta)}{\partial \theta \partial \theta'} \right]_{\theta = \theta^*}^{-1}.
  \]
Metropolis-Hastings Algorithm (MCMC) ...

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

  - $\theta^{(1)} = \theta^\ast$

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

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

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

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

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

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

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

• Some intuition

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

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

  – Set, \( \{ \theta^{(1)}, \theta^{(2)}, ..., \theta^{(M)} \} \), 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

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

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

$$f(\theta) \sim \text{prior on parameters}$$

$$\theta^* \sim \text{maximum of } g(\theta) \ (\text{i.e., mode})$$
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:

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

- 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.