Approximations based on Posterior Modes

Statistics 220

Spring 2005

Posterior Modes

As we have seen earlier, often as $n \to \infty$

$$p(\theta|y) \approx N(\hat{\theta}, [I(\hat{\theta})]^{-1})$$

where $\hat{\theta}$ is the the posterior mode and $I(\theta)$ is the observed information

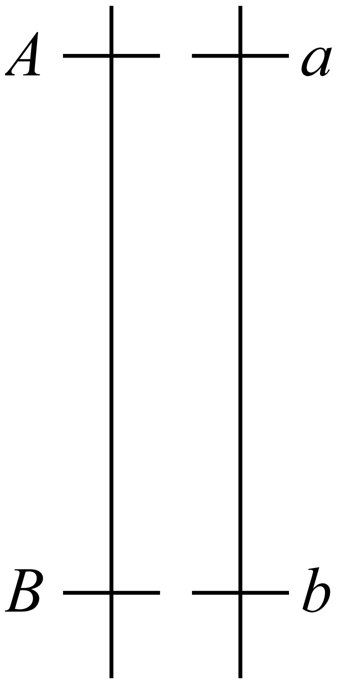
$$I(\theta) = -\frac{d^2}{d\theta^2} \log p(\theta|y)$$

While the asymptotics are interesting, the approximate normality is helpful in other situations.

1. Crude estimates as starting points for approximations
2. Normal (or related) mixture approximations to the posterior
3. Separate approximations for different marginal and conditional posterior distributions
4. Approximating distributions for Monte Carlo methods (e.g. a proposal in Metropolis-Hastings or in Importance Sampling)

To find the posterior mode and information, numerical methods often need to be used as closed form solutions usually aren't available.

Example: Linkage Analysis (Rao, 1973, pp 268-269)



Two genes on a chromosome are separated by a recombination fraction $\theta \leq \frac{1}{2}$

For an organism with joint haplotype $AB|ab$, there are 4 possible haplotypes that can be passed to its offspring

Haplotype	Probability
AB	$\frac{1-\theta}{2}$
Ab	$\frac{\theta}{2}$
aB	$\frac{\theta}{2}$
ab	$\frac{1-\theta}{2}$

An experiment was performed to estimate θ. The breeding experiment used $AB|ab \times AB|ab$ crosses and recorded the observed phenotypes. In this experiment, 2 dominant traits were observed (A dominant to a and B dominant to b).

While there are 16 different possible joint haplotypes in the offspring (4 from the father times 4 from the mother), there are only 4 possible phenotypes.

Phenotype	Probability	Counts
AB	$\frac{3-2\theta+\theta^2}{4}$	125
Ab	$\frac{2\theta-\theta^2}{4}$	18
aB	$\frac{2\theta-\theta^2}{4}$	20
ab	$\frac{1-2\theta+\theta^2}{4}$	34

So the likelihood function is

$$p(y|\theta) = (3 - 2\theta + \theta^2)^{125}(2\theta - \theta^2)^{18+20}(1 - 2\theta + \theta^2)^{34}$$

Now lets put a truncated Beta prior on θ

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} I(\theta \leq 0.5)$$

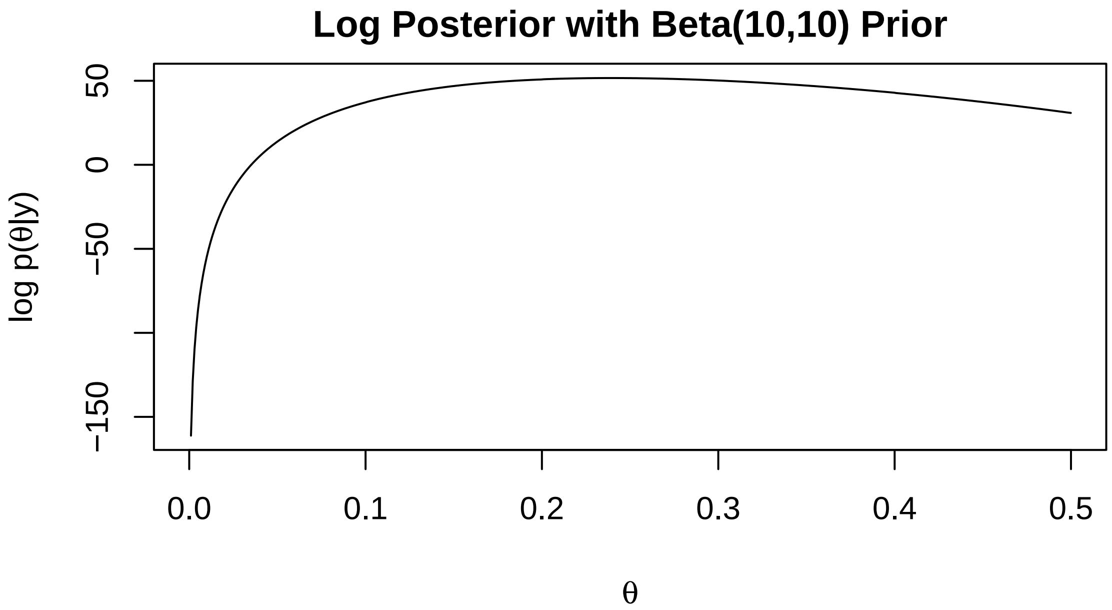
Thus the log posterior is

$$\begin{aligned} \log p(\theta|y) &= 125\log(3-2\theta+\theta^2) + 38\log(2\theta-\theta^2) + 34\log(1-2\theta+\theta^2) \\ &\quad +(\alpha-1)\log\theta + (\beta-1)\log(1-\theta) \\ &= 125\log(3-2\theta+\theta^2) + 38\log(2-\theta) + \\ &\quad (\alpha+37)\log\theta + (\beta+67)\log(1-\theta) \end{aligned}$$

Following the usual approach solving $\frac{d}{d\theta}\log p(\theta|y) = 0$ to optimize gives

$$\frac{d}{d\theta}\log p(\theta|y) = \frac{125(2\theta-2)}{3-2\theta+\theta^2} - \frac{38}{2-\theta} + \frac{\alpha+37}{\theta} - \frac{\beta+67}{1-\theta} = 0$$

This does not have an obvious closed form solution so we need to find another approach to maximizing the posterior (or log posterior).



There are a wide array of numerical approaches for optimizing functions. I want to discuss two

1. Newton-Raphson (and approximations)

2. EM algorithm

Newton-Raphson

Following the text, let

$$L(\theta) = \log p(\theta|y)$$

Note that this can be an unnormalized density as

$$L_c(\theta) = \log cp(\theta|y) = L(\theta) + \log c$$

as the same optima (c can't be a function of θ). So we can also use

$$L(\theta) = \log p(y|\theta)p(\theta) = \log p(y|\theta) + \log p(\theta)$$

as the function to optimize.

So we want to solve the function

$$L'(\theta) = 0$$

where $L'(\theta)$ is the vector of first partial derivatives (i.e. the gradient).

For Newton-Raphson, we also need $L''(\theta)$, the matrix of second partial derviatives.

The the Newton-Raphson algorithm is

1. Choose a starting value, θ^0
2. For $t = 1, 2, 3, \ldots$
   (a) Compute $L'(\theta^{t-1})$ and $L''(\theta^{t-1})$. The Newton method step at time t is based on the quadratic approximation to $L(\theta)$ centered at θ^{t-1}.
   (b) Set the new iterate, θ^t, to maximize the quadratic approximation

$$\theta^t = \theta^{t-1} - [L''(\theta^{t-1})]^{-1} L'(\theta^{t-1})$$

So for the example

$$L'(\theta) = \frac{250(\theta - 1)}{3 - 2\theta + \theta^2} - \frac{38}{2 - \theta} + \frac{\alpha + 37}{\theta} - \frac{\beta + 67}{1 - \theta}$$

$$L''(\theta) = \frac{250}{(3 - 2\theta + \theta^2)} - \frac{500(\theta - 1)^2}{(3 - 2\theta + \theta^2)^2} - \frac{38}{(2 - \theta)^2} - \frac{\alpha + 37}{\theta^2} - \frac{\beta + 67}{(1 - \theta)^2}$$

Note that Newton-Raphson is not guaranteed to converge. The starting point θ^0 can be very important, particularly when $-L''$ is not positive definite.

One advantage to Newton-Raphson is that once you get close to the solution, the convergence is very fast (quadratic convergence). Also if the sequence won't converge, it is usually obvious quickly.

There is another advantage to Newton-Raphson in the Bayesian (or likelihood) framework. The update formula can be written as

$$\theta^t = \theta^{t-1} - [L''(\theta^{t-1})]^{-1} L'(\theta^{t-1}) = \theta^{t-1} + [I(\theta^{t-1})]^{-1} L'(\theta^{t-1})$$

Thus as we are determining the mode, we are also calculating the information matrix, and depending on how the update is done, we are also getting the asymptotic posterior variance matrix $[I(\theta)]^{-1}$.

Note that when implementing this, you usually do not want to invert $I(\theta^{t-1})$, but instead solve the system

$$L''(\theta^{t-1})\Delta\theta = L'(\theta^{t-1}) \quad \text{or} \quad I(\theta^{t-1})\Delta\theta^* = L'(\theta^{t-1})$$

and update with

$$\theta^t = \theta^{t-1} - \Delta\theta \quad \text{or} \quad \theta^t = \theta^{t-1} + \Delta\theta^*$$

as it is faster and more numerically stable.

Approximations to Newton-Raphson

Note as described, Newton-Raphson requires the calculations of derivatives. However it is easy to approximate derivatives numerically. One approach is approximate the derivatives with

$$L_i'(\theta) = \frac{dL}{d\theta_i} \approx \frac{L(\theta + \delta_i e_i) - L(\theta - \delta_i e_i)}{2\delta_i}$$

and

$$\begin{aligned}
L_{ij}''(\theta) = \frac{d^2 L}{d\theta_i d\theta_j} &= \frac{d}{d\theta_j}\frac{dL}{d\theta_i} \\
&\approx \frac{L_i'(\theta + \delta_j e_j) - L_i'(\theta - \delta_j e_j)}{2\delta_j} \\
&\approx [L(\theta + \delta_i e_i + \delta_j e_j) - L(\theta - \delta_i e_i + \delta_j e_j) \\
&\qquad -L(\theta - \delta_i e_i + \delta_j e_j) + L(\theta - \delta_i e_i - \delta_j e_j)]/(4\delta_i\delta_j)
\end{aligned}$$

where e_i is the unit vector corresponding to the ith component of θ.

δ_i, the size of the deviation to take along direction e_i depends on the scale of the problem, but should be small.

You don't want it too big as curvature of L will make this a poor approximation.

But you don't want it too small as to avoid round off error.

Often a value such as 0.0001 is reasonable.

EM Algorithm

Dempster, Laird, and Rubin (1977)

EM $=$ Expectation – Maximization

An approach for finding MLEs and posterior modes.

In the likelihood situation, it is often based on decomposing data $X = (Y, Z)$ into observed (Y) and missing parts (Z). Want to maximize

$$p(y|\theta) = \int p(y, z|\theta) dz$$

In the Bayesian situation, its based on splitting $\theta = (\phi, \gamma)$, where you want to maximize over ϕ after average over γ.

$$p(\phi|y) = \int p(\phi, \gamma|y) d\gamma$$

Want posterior mode of $p(\phi|y)$ instead of $p(\phi, \gamma|y)$.

I will present thing in terms of the Bayesian solution. For the implementation in the likelihood situation, see the 221 notes on the course web site.

EM in this setting is based on the relationship

$$p(\phi|y) = \frac{p(\phi, \gamma|y)}{p(\gamma|\phi, y)}$$

Now lets take logs, giving

$$\log p(\phi|y) = \log p(\phi, \gamma|y) - \log p(\gamma|\phi, y)$$

Lets take expectation of both sides, with respect to the density $p(\gamma|\phi^{\text{old}}, y)$, where ϕ^{old} is a current (old) guess of ϕ

$$\log p(\phi|y) = E_{\text{old}}[\log p(\phi, \gamma|y)] - E_{\text{old}}[\log p(\gamma|\phi, y)]$$

Let

$$Q(\phi|\phi^{\text{old}}) = E_{\text{old}}[\log p(\phi, \gamma|y)]$$

and

$$H(\phi|\phi^{\text{old}}) = E_{\text{old}}[\log p(\gamma|\phi, y)]$$

So

$$\log p(\phi|y) = Q(\phi|\phi^{\text{old}}) - H(\phi|\phi^{\text{old}})$$

It is possible to show that $H(\phi|\phi^{\text{old}})$, treated as a function of ϕ, is maxmized at ϕ^{old}, i.e.

$$H(\phi|\phi^{\text{old}}) \leq H(\phi^{\text{old}}|\phi^{\text{old}}) \quad \text{for all } \phi$$

Now let ϕ^{new} be any value of ϕ such that

$$Q(\phi^{\text{new}}|\phi^{\text{old}}) \geq Q(\phi^{\text{old}}|\phi^{\text{old}})$$

Thus

$$\begin{aligned} \log p(\phi^{\text{new}}|y) &= Q(\phi^{\text{new}}|\phi^{\text{old}}) - H(\phi^{\text{new}}|\phi^{\text{old}}) \\ &\geq Q(\phi^{\text{old}}|\phi^{\text{old}}) - H(\phi^{\text{old}}|\phi^{\text{old}}) = \log p(\phi^{\text{old}}|y) \end{aligned}$$

This relationship is the main idea behind the Generalized EM (GEM) algorithm.

At each step, finding a ϕ which leads to an increase of $Q(\phi|\phi^{\text{old}})$ must lead to an increase in the marginal log posterior $\log p(\phi|y)$.

Implementing the EM Algorithm

1. Start with a estimate of the parameter ϕ^0.

2. For $t = 1, 2, 3, \ldots$

   (a) E-step: Determine the expected log posterior density function

   $$Q(\phi|\phi^{t-1}) = E_t[\log p(\phi, \gamma|y)] = \int \log p(\phi, \gamma|y) p(\gamma|\phi^{t-1}, y) d\gamma$$

   (b) M-step: Maximize the expected log posterior density

   $$\phi^t = \arg\sup Q(\phi|\phi^{t-1})$$

   For a GEM algorithm, ϕ^{t-1} doesn't have to maximize $Q(\phi|\phi^{t-1})$ but only satisfy $Q(\phi^t|\phi^{t-1}) \geq Q(\phi^{t-1}|\phi^{t-1})$.

Example: Fatal Airline Accidents

$$\begin{aligned} y_i &\overset{iid}{\sim} Poisson(\lambda) \\ \lambda &\sim Exp(\mu) \\ \mu &\sim Gamma(\alpha, \beta) \end{aligned}$$

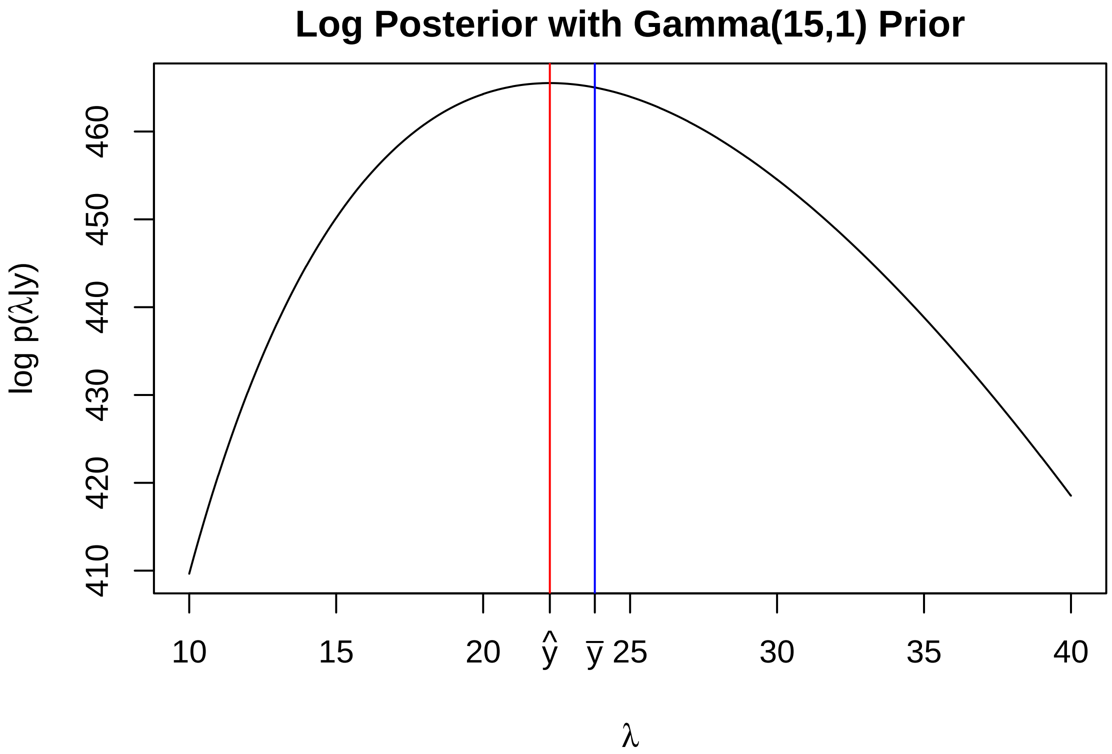
Want to maximize $p(\lambda|y)$. Note that

$$p(\lambda|y) \propto \frac{\lambda^{n\bar{y}} e^{-n\lambda}}{(\lambda+\beta)^{\alpha+1}}$$

The mode is a solution to the quadratic equation

$$n\lambda^2 + (\alpha + 1 + n\beta - n\bar{y})\lambda - n\beta = 0$$

For the data in Table 2.2, $n = 10$ and $\bar{y} = 23.8$. If $\alpha = 15$ and $\beta = 1$



Need for EM algorithm:

1. $p(\lambda, \mu|y)$

$$p(\lambda, \mu|y) \propto \lambda^{\sum y_i} e^{-n\lambda} \mu e^{-\mu\lambda} \mu^{\alpha-1} \beta^{\alpha} \frac{e^{-\mu\beta}}{\Gamma(\alpha)}$$

   Note that the sufficient statistic here is $\bar{y} = 23.8$ and $n = 10$

2. $p(\mu|\lambda, y)$

$$\mu|\lambda, y \sim Gamma(\alpha + 1, \lambda + \beta)$$

3. E-step: find $Q(\lambda|\lambda^{t-1})$

$$\log p(\lambda, \mu|y) = \alpha \log \mu - \mu(\lambda + \beta) - n\lambda + n\bar{y} \log \lambda + c$$

$$\begin{aligned} Q(\lambda|\lambda^{t-1}) &= E_{\lambda^{t-1}}[\mu\lambda - n\lambda + n\bar{y} \log \lambda + c] \\ &= \lambda E_{\lambda^{t-1}}[\mu] + n\lambda + n\bar{y} \log \lambda + c \\ &= \lambda(E_{\lambda^{t-1}}[\mu] + n) + n\bar{y} \log \lambda + c \end{aligned}$$

since $\mu|\lambda, y$ is $Gamma(\alpha + 1, \lambda + \beta)$

$$E_{\lambda^{t-1}}[\mu] = \frac{\alpha + 1}{\lambda^{t-1} + \beta}$$

then

$$Q(\lambda|\lambda^{t-1}) = \lambda \left(\frac{\alpha + 1}{\lambda^{t-1} + \beta} + n \right) + n\bar{y} \log \lambda + c$$

4. M-step:

$$\begin{aligned}\lambda^t &= \arg\sup Q(\lambda|\lambda^{t-1}) \\ &= \frac{n\bar{y}}{\frac{\alpha+1}{\lambda^{t-1}+\beta}+n} \\ &= \frac{n\bar{y}(\lambda^{t-1}+\beta)}{\alpha+1+n(\lambda^{t-1}+\beta)}\end{aligned}$$

Note that it can be shown that this sequence will converge to a root of

$$n\lambda^2 + (\alpha+1+n\beta-n\bar{y})\lambda - n\beta = 0$$

which is the same equation derived for $\log p(\lambda|y)$ directly.

If $\alpha = 15$ and $\beta = 1$ and $\lambda^0 = 15$, the sequence of updates is

t	λ^t
0	15.00000
1	21.63636
2	22.22881
3	22.26630
4	22.26861
5	22.26875
6	22.26876
7	22.26876

Under some regularity conditions, for any GEM, the sequence $\phi^1, \phi^2, \phi^3, \ldots$ converges to a local mode of the posterior density.

Note that the proof of this result in Dempster, Laird, and Rubin (1979) wasn't quite right. Wu (1983) found valid conditions to indicate when this sequence would converge to a local mode.

Theorem. *Under some regularity conditions, for any GEM sequence* $\{\phi^t\}$,

$$\log p(\phi^t|y) > \log p(\phi^{t-1}|y)$$

if

$$\phi^{t-1} \notin \Phi = \left\{ \phi : \frac{d}{d\phi} \log p(\phi|y) = 0 \right\}$$

Thus you will continue to increase the posterior density until you hit a local mode.

The proof of the theorem depends on the fact that

$$\frac{d}{d\phi} \log p(\phi|y) = \frac{d}{d\phi} Q(\phi|\phi)$$

when the derivative of Q is taken with respect to the first ϕ. Thus if you are at a mode, Q must have a 0 derivative, implying you can't increase Q.

The EM algorithm has linear convergence, thus it tends to converge slower than algorithms like Newton-Raphson. However it does have the advantage that it is guaranteed to converge, unlike Newton-Raphson.

For this algorithm to be feasible, it must be possible to maximize Q easily, or at least find a value which increases it. Thus the EM algorithm isn't feasible for all problems. However there are a number of extensions that can make some of the more difficult problems feasible.