

Chapter 5: Advanced computations

5.1: Approximations based on Posterior modes:

We have

$p(\theta|y)$: Target distribution

if θ is a high dimensional parameter, we can factorize it as

$$\begin{aligned} p(\theta|y) &= p(v, \phi|y) \\ &= p(v|\phi, y)p(\phi|y) \end{aligned}$$

and compute the conditional and the marginal posterior densities.

5.2: Normalized and unnormalized densities:

We suppose that $P(\theta|y)$ is easily computed up to a proportionality constant

It means that it exists $q(\theta|y)$, an unnormalized density, for which

$$\frac{q(\theta|y)}{p(\theta|y)} = \text{const (depend on } y \text{ only)}$$

Trick: is to use the $\log p(\theta|y)$ or $q(\theta|y)$ to avoid overflow and underflow.

Rough estimate of the location of the mode/modes:

Approximation of the posterior density if it is multimodal, an approximation must be provided near each mode.

- In general, if we have many modes, we find them
- We search in general for a single mode and if it does not look reasonable, we continue searching through the parameter space for other modes.

There we will introduce two simple methods that are commonly used for solving optimizing problems to find modes

(a) Conditional maximization (CM)

Called: stepwise ascent:

- 1.** start with a rough estimate of θ
- 2.** alter one component of θ at a time, leaving the other components at their previous values
- 3.** at each step increase the $\log p(\theta|y)$ (if it is bounded).
- 4.** convergence to local mode.

If the posterior distribution has a simple analytic form and is easy to be maximized. In this case, just maximize the density with respect to one parameter at a time, iterating until the steps become small enough that approximate convergence has been reached.

(b)Newton's method

Acceptable for unnormalized posterior

Called also: Newton-Raphson algorithm: based on a quadratic Taylor series approximation of the $\log p(\theta|y)$

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

- determine $L'(\theta)$ and $L''(\theta)$, vectors of derivatives and matrix of second derivatives
- choose a starting value, θ^0
- for $t = 1, 2, \dots$
 - compute $L'(\theta^{t-1})$ and $L''(\theta^{t-1})$. The newton's method step at time t is based on the quadratic approximation of $L(\theta)$ centered at θ^{t-1} .
 - set the new iterate, θ^t , to maximize the quadratic approximation; thus

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

Starting values are important.

Starting values may be obtained from crude estimation, or conditional maximization could be used to generate a starting value for Newton's method.

Remark: convergence is very fast once the iterates are close to the solution.

If the derivatives (first and second) are difficult to determine analytically, one can approximate them numerically using finite differences.

Each component of L' can be estimated numerically at any specified value $\theta = (\theta_1, \dots, \theta_p)$ by

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

where δ_i is a small value and, using linear algebra notation,

e_i is the unit vector corresponding to the i th component of θ .

δ_i are chosen based on the scale of the problem,

example: $\delta = 0.0001$

and

$$L''_{ij}(\theta) = \frac{d^2 L}{d\theta_i d\theta_j} = \frac{d}{d\theta_j} \left(\frac{dL}{d\theta_i} \right) \approx \frac{L'_i(\theta + \delta_j e_j | y) - L'_i(\theta - \delta_j e_j | y)}{2\delta_j}$$

$$\approx [L(\theta + \delta_i e_i + \delta_j e_j) - L(\theta + \delta_i e_i - \delta_j e_j) - L(\theta - \delta_i e_i + \delta_j e_j) + L(\theta - \delta_i e_i - \delta_j e_j)] / (4\delta_i \delta_j)$$

5.4: The normal and related mixtures approximations:

When the modes are found, we can construct an approximations based on the multivariate normal distribution.

Unimodal density:

$$P_{normal\ approx}(\theta) = N(\theta | \hat{\theta}, V_\theta)$$

(fit a normal distribution to the first two derivatives of the log $p(\theta|y)$ at the mode $\hat{\theta}$.)

$$V_\theta = [L''(\hat{\theta})]^{-1}$$

5.4.1: Multimodal densities:

If we have found K modes in the posterior density, the posterior distribution can then be approximated by a mixture of K multivariate normals, each with its own mode $\hat{\theta}_k$, variance matrix V_{θ_k} and mixing proportion ω_k

$$P_{normal\ approx}(\theta) \approx \sum_{k=1}^K \omega_k N(\theta | \hat{\theta}_k, V_{\theta_k})$$

ω_k is approximated by comparing $p(\hat{\theta}_k|y)$ to the approximate density, $P_{normal\ approx}(\hat{\theta})$ at each of the K modes.

If the modes are fairly widely separated and the normal approximation is appropriate for each mode, then we obtain

$$\omega_k = (p(\hat{\theta}_k|y) |V_{\theta_k}|)^{1/2}$$

which yields the normal-mixture approximation

$$P_{normal\ approx}(\theta) \propto \sum_{k=1}^K p(\hat{\theta}_k|y) \exp\left(-\frac{1}{2}(\theta - \hat{\theta}_k)^t V_{\theta_k}^{-1}(\theta - \hat{\theta}_k)\right)$$

5.4.2: t -approximation:

For small samples, it is useful for the initial approximating distribution to cover more of the parameter space to ensure that nothing is missing.

We then replace the normal approximation by the multivariate *Student_t* with small number of degrees of freedom, ν .

$$P_{student-t\ approx}(\theta) \propto \sum_{k=1}^K p(\hat{\theta}_k|y) \exp\left(\nu + (\theta - \hat{\theta}_k)^t V_{\theta_k}^{-1}(\theta - \hat{\theta}_k)\right)^{-(p+\nu)/2}$$

Several different strategies can be employed to improve the approximate distribution further:

- analytically fitting the approximation to locations other than the modes, such as saddle point...
- analytically or numerically integrating out some components of the

target distribution to obtain a lower-dimensional approximation

- bounding the range of parameter values.

5.5 Finding marginal posterior modes using EM and related

when there are many parameters, normal approximations to the joint is useless and the joint is not helpful.

So we use the marginal posterior mode of a subset of parameters

let's define

$$\theta = (\gamma, \phi)$$

and let's suppose that we are interested in approximating

$$p(\phi|y)$$

- use the normal or t mixture distribution for $p(\phi|y)$
- approximate $p(\gamma|\phi, y)$ as a normal or t mixture with point depending on ϕ .

If

$$p(\phi|y)$$

is defined analytically, we can use one of the previous optimization methods.

5.5.1 EM algorithm:

EM algorithm is one iterative process for finding the **mode** of the marginal posterior density $p(\phi|y)$ where:

- ϕ will be the parameter to estimate
- γ is the missing data.

Aim of the EM:

Is handling missing data as the following:

- 1.** replace missing data by their expectations given the parameters
- 2.** Estimate parameters assuming the missing data are given by their estimated values
- 3.** Reestimate the missing values assuming the new parameter estimates are correct.
- 4.** Reestimate parameter and so forth until convergence

EM: stands for Expectation + Maximization

- 1.** Expectation: by finding the expectation of the needed functions of the missing values
- 2.** Maximizing: by estimating the parameters as if these functions of the missing data were observed.

EM finds the mode of the marginal posterior distribution

$$p(\phi|y)$$

averaging over the parameter γ

Each iteration of the EM algorithm increases

$$\log p(\phi|y)$$

until convergence.

5.5.2: Derivation of the EM algorithm and Why it works?

$$p(\phi|y) = p(\gamma, \phi|y) / p(\gamma|\phi, y)$$

this implies that:

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

integrate out γ with respect to its distribution

$$\begin{aligned} E(\log(\phi|y)) &= E(\log p(\gamma, \phi|y)) - E(\log p(\gamma|\phi, y)) \\ &= \int (\log p(\gamma, \phi|y)) p(\gamma|\phi, y) d\gamma - \int \log p(\gamma|\phi, y) p(\gamma|\phi, y) d\gamma \end{aligned}$$

$$Q(\phi, \phi^*) = \int (\log p(\gamma, \phi|y)) p(\gamma|\phi^*, y) d\gamma$$

and

$$H(\phi, \phi^*) = \int \log p(\gamma|\phi, y) p(\gamma|\phi^*, y) d\gamma$$

$$\begin{aligned} &\log p(\phi^{(i+1)}|y) - \log p(\phi^{(i)}|y) \\ &= Q(\phi^{(i+1)}, \phi^{(i)}) - Q(\phi^{(i)}, \phi^{(i)}) \\ &\quad + H(\phi^{(i+1)}, \phi^{(i)}) - H(\phi^{(i)}, \phi^{(i)}) \end{aligned}$$

we choose $\phi^{(i+1)}$ so that

$$Q(\phi^{(i+1)}, \phi^{(i)}) \geq Q(\phi^{(i)}, \phi^{(i)})$$

Dealing with H

$$\begin{aligned}
& H(\phi^{(i+1)}, \phi^{(i)}) - H(\phi^{(i)}, \phi^{(i)}) \\
&= \int p(\gamma|\phi^{(i)}, y) \log \frac{p(\gamma|\phi^{(i+1)}, y)}{p(\gamma|\phi^{(i)}, y)} d\gamma \\
&= E_{\phi^{(i)}} \left[\log \frac{p(\gamma|\phi^{(i+1)}, y)}{p(\gamma|\phi^{(i)}, y)} \right] \\
&\geq \log E_{\phi^{(i)}} \left[\frac{p(\gamma|\phi^{(i+1)}, y)}{p(\gamma|\phi^{(i)}, y)} \right] \text{Jensen's inequality} \\
&= 0
\end{aligned}$$

- So the Q terms and the H terms are non-negative.
- So the steps do not decrease the log-likelihood

5.6 Application

Normal data with *Unknown mean and variance* (μ, σ^2) using semi-conjugate prior

$$y_1, \dots, y_n \sim N(\mu, \sigma^2)$$

$$\mu \sim N(\mu_0, \tau_0^2), \mu_0$$

$$\log(\sigma) \propto 1 \text{ (non-informative prior on } \sigma^2)$$

Using Bayesian approach, it will be hard to derive the marginal posterior distribution $p(\mu|y)$.

Here we can use the EM Algorithm to find the marginal posterior mode of μ averaging over σ . So we have here $(\phi, \gamma) = (\mu, \sigma)$.

The joint log posterior density is:

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

$$= -\frac{1}{2\tau_0^2}(\mu - \mu_0) + (n + 1)\log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 + C_1$$

E-step:

Determine the expected log posterior density function

$$E \log p(\mu, \sigma | y) = -\frac{1}{2\tau_0^2}(\mu - \mu_0)^2 + (n + 1)E(\log \sigma) - \frac{1}{2}E\left(\frac{1}{\sigma^2}\right) \sum_{i=1}^n (y_i - \mu)^2$$

Page 46 and 47 of the book or chapter 2, we have:

Normal data with known mean and unknown variance:

$$p(\sigma^2) \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

$$(y | \sigma^2) \sim N(\mu, \sigma^2)$$

$$(\sigma^2 | y) \sim IG\left(\frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + n \sum (y_i - \mu)^2}{2}\right)$$

here when

$$\sigma^2 \sim \text{non informative prior, then } \nu_0 = 0$$

so

$$(\sigma^2 | y) \sim IG\left(\frac{n}{2}, \frac{n \sum (y_i - \mu)^2}{2}\right)$$

and then

$$(1/\sigma^2 | y) \sim G\left(\frac{n}{2}, \frac{n \sum (y_i - \mu)^2}{2}\right)$$

and then

$$E(1/\sigma^2) = \left(\frac{\sum (y_i - \mu)^2}{n} \right)^{-1}$$

so we have

$$E \log p(\mu, \sigma | y) = -\frac{1}{2\tau_0^2} (\mu - \mu_0)^2 - \frac{1}{2} \left(\frac{\sum (y_i - \mu_{old})^2}{n} \right)^{-1} \sum_{i=1}^n (y_i - \mu)^2 + C_2$$

Here the term $E(\log \sigma)$ does not depend on μ ; so the maximization step for μ is not affected by $E(\log \sigma)$.

M step:

Here we must find ϕ (means μ) that maximizes $E(\log p(\mu, \sigma | y))$

Remark,

$$\begin{aligned} & -\frac{1}{2\tau_0^2} (\mu - \mu_0)^2 - \frac{1}{2} \left(\frac{\sum (y_i - \mu_{old})^2}{n} \right)^{-1} \sum_{i=1}^n (y_i - \mu)^2 + C_2 \\ \propto & \log \left(N(\mu | \mu_0, \tau_0^2) \cdot \prod_{i=1}^n p(y_i | \mu, \frac{\sum_{i=1}^n (y_i - \mu_{old})^2}{n}) \right) \\ = & \log \left(p(\mu) \cdot p(y | \mu, \sigma^2 = \frac{\sum_{i=1}^n (y_i - \mu_{old})^2}{n}) \right) \\ = & \log[p(\mu | y)] \end{aligned}$$

which is the case of a normally distributed data with *unknown* mean μ and *known* variance $\sigma^2 = \frac{\sum_{i=1}^n (y_i - \mu_{old})^2}{n}$

and then the maximum (mode) is obtained at

$$\begin{aligned}
\hat{\mu}_{new} &= \frac{\frac{1}{\tau_0^2} \mu_0 + \frac{n}{\sigma^2} \bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \\
&= \frac{\frac{1}{\tau_0^2} \mu_0 + \frac{n}{\left(\frac{\sum_{i=1}^n (y_i - \mu_{old})^2}{n}\right)} \bar{y}}{\frac{1}{\tau_0^2} + \left(\frac{n}{\sum_{i=1}^n (y_i - \mu_{old})^2}\right)}
\end{aligned}$$

we iterate these two steps until convergence to the marginal posterior mode $p(\mu|y)$.