STAT 406 Lab 11 Study guide and hint

In the lab(11:30am), we covered the concept of **loglikelihood function** $\ell(\theta|data)$ and **maximum likelihood estimator(mle)** $\hat{\theta}$, where *data* are observations x_i of random variables X_i with densities $\pi_i(x;\theta)$. To get the **mle** $\hat{\theta}$, we need to maximize $\ell(\theta|data)$ or minimize $-\ell(\theta|data)$ over θ, so we introduced **gradient method and Newton Ralphson method** as two basic algorithm to optimize a general nice function like $-\ell(\theta|data)$ (those we have seen in the class). R base function $optim()$ is a choice for you to optimize a function in a software blackbox.

The **EM algorithm** is useful when the model is a latent variable model or contain missing data, where it's not practical to optimize the generally intractable **loglikelihood function** $\ell(\theta|data)$ (we can derive the loglikelihood function though - read in the end), so we work with a smart surrogate function, the **"complete" loglikelihood function** $\ell_c(\theta|data, unobserved\ variables)$ which is more tractable. However, the function $\ell_c()$ is a function of both θ and *unobserved variables*, to reduce it in a statistically sensible way to a function of θ alone that we can maximize over ~~and get the $\hat{\theta}$~~, we derive the conditional average $\mathbb{E}[\ell_c(\theta|data, unobserved\ variables)\,|data]$ as a function of θ alone, maximize it, and do the conditional expectation and maximization iteratively until convergence to get $\hat{\theta}$, where the expectation is taken over all the *unobserved variables* condtioning on the *data*.

In general in this class,

Given data $\mathbf{x} = \{x_1, \ldots, x_n\}$ as observations of independent random variables $\{X_1, \ldots, X_n\}$ with marginal densities $X_i \sim \pi_i(x;\theta)$, $i = 1, \ldots, n$, loglikelihood function is given by

$$\ell(\theta|\mathbf{x}) = \sum_{i=1}^{n} \log(\pi_i(x_i;\theta)) \tag{0.0.1}$$

We then optimize $\ell(\theta|\mathbf{x})$ or $-\ell(\theta|\mathbf{x})$, using the gradient method or Newton Ralphson method, or $optim()$ in R to get $\hat{\theta}$.

A Study Example: lecture note "Handout2Chap4-2.pdf" pages 35 - 39, where the NRMethod function on page 38 gives a general purpose NR implementation, but you need to code the input arguments "fun, grad, hess" according to your problem at hand.

In the case of latent variable models where EM algorithm is used, the model specifies the conditional densities of random variables X_i's conditioning on the latent variables Z_i's as $\pi_i(x_i|z_i;\theta)$ independently over $i = 1, \ldots, n$. The model also specifies the marginal densities of independent Z_i's as $Z_i \sim p(z;\vartheta)$. Now we can derive the

joint densities of X_i and Z_i's as $q_i(x_i, z_i; \theta, \vartheta) = \pi_i(x_i|z_i;\theta)\, p(z_i;\vartheta)$, for $i = 1:n$. And the **complete loglikelihood function** is given by

$$\begin{aligned}\ell(\theta, \vartheta|\mathbf{x}, \mathbf{z}) &= \sum_{i=1}^{n} \log\left(q_i(x_i, z_i; \theta, \vartheta)\right) \\ &= \sum_{i=1}^{n} \log\left(\pi_i(x_i|z_i;\theta)\, p(z_i;\vartheta)\right) \\ &= \sum_{i=1}^{n} \left(\log\left(\pi_i(x_i|z_i;\theta)\right) + \log\left(p(z_i;\vartheta)\right)\right)\end{aligned}$$

The EM algorithm does the following: suppose we initialize parameters at θ_0, ϑ_0, for $k = 0, 1, \ldots, K$, iterate until *convergence*(choose a relatively large K, loop until two consecutive updates differ less than a small tolerance value):

1. E-step : set $Q(\theta, \vartheta|\theta_k, \vartheta_k) = \mathbb{E}\left[\ell(\theta, \vartheta|\mathbf{x}, \mathbf{z})\,|\mathbf{x}, \theta_k, \vartheta_k\right]$.

   The above notation means given the *data* $\mathbf{x}$ and last update values θ_k, ϑ_k, you will use the conditional probability density/mass $p(z|x, \theta_k, \vartheta_k)$ to get the expectation of the random quantity $\ell(\theta, \vartheta|\mathbf{x}, \mathbf{z})$ which is a function of random variable $\mathbf{z}$. The θ, ϑ are unknown parameters ;

2. M-step: maximize $Q(\theta, \vartheta|\theta_k, \vartheta_k)$ over (θ, ϑ), and update $(\theta_{k+1}, \vartheta_{k+1}) = \text{argmax}(Q(\theta, \vartheta|\theta_k, \vartheta_k))$.

3. Substitute the updated $(\theta_{k+1}, \vartheta_{k+1})$ for (θ_k, ϑ_k) in E-step, (continue with the E/M steps).

A Study Example: Lecture notes "HandoutChap5" page 35-37, write the codes, study the plots. Lab notes "Lab_11.pdf" 3.2 1-dimensional Gaussian mixture. Notice the trick that in general for a bernoulli random variable $w \sim \text{Ber}(p)$, with $p := \mathcal{P}(w = 1)$, its pmf can be conveniently expressed as $\pi(w) = p^w(1-p)^{(1-w)}$. Notice also how the **Bayes Formula** is used to compute $p(Z_i = 1|X_i) = \mathbb{E}(Z_i|X_i) = 1 \cdot p(Z_i = 1|X_i) + 0 \cdot p(Z_i = 0|X_i)$.

We can also derive the marginal densities of X_i's as

- in the case of discrete Z's (lab):

$$\pi_i(x_i;\theta) = \sum_{j=1}^{n} \pi_i(x_i|z_j;\theta)\, p(z_j;\vartheta)\,;$$

- in the case of continuous Z's (homework):

$$\pi_i(x_i;\theta) = \int_{z_j} \pi_i(x_i|z_j;\theta)\, p(z_j;\vartheta)\, dz_j;$$

So the **loglikelihood function** is given by (recall (0.0.1))

$$\begin{aligned}\ell(\theta|\mathbf{x}) &= \sum_{i=1}^{n} \log\left(\pi_i\left(x_i;\theta\right)\right) \\ &= \begin{cases} \sum_{i=1}^{n} \log\left(\sum_{j=1}^{n} \pi_i\left(x_i|z_j;\theta\right) p\left(z_j;\vartheta\right)\right), & \text{discrete } z_j's \\ \sum_{i=1}^{n} \log\left(\int_{z_j} \pi_i\left(x_i|z_j;\theta\right) p\left(z_j;\vartheta\right) dz_j\right), & \text{continuous } z_j's \end{cases}\end{aligned}$$

which is generally intractable due to the presence of logarithm of sum or integration.

Homework 10 hint: In homework 10(2.a) however, the function $\pi_i\left(x_i|z_j;\theta\right)$ does **not** involve z_j as a variable, so the marginal density of X_i can be reduced as $\pi_i\left(x_i;\theta\right) = \int_{z_j} \pi_i\left(x_i|z_j;\theta\right) p\left(z_j;\vartheta\right) dz_j = \pi_i\left(x_i|z_j;\theta\right) \int_{z_j} p\left(z_j;\vartheta\right) dz_j = \pi_i\left(x_i|z_j;\theta\right)$. where integration of density of Z_j gives 1. Suddenly the loglikelihood function $\ell(\theta|\mathbf{x})$ becomes tractable and can be maximized directly! - not the case in general. Notice the problem also asks you to apply EM algorithm on complete loglikelihood function and compare the EM solution with the solution you get from maxmizing the loglikelihood function $\ell(\theta|\mathbf{x}) = \log(L(\theta))$ directly. *optim*() in general minimize the input function, and maximize $\log(L(\theta))$ is the same as minimize $-\log(L(\theta))$.