STAT3500/7500

Assignment 3

(a) [5 marks]

Let $\boldsymbol{x} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T$ contain the observed values of a random sample

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} F_{\boldsymbol{\theta}}, \tag{1}$$

where $F_{\boldsymbol{\theta}}$ denotes a distribution function indexed by the parameter vector $\boldsymbol{\theta}$.

Suppose that the likelihood function $L(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ belongs to the regular exponential family so that it is of the form

$$L(\theta) = b(\boldsymbol{x}) \exp\{\boldsymbol{c}(\boldsymbol{\theta})^T \boldsymbol{t}(\boldsymbol{x})\}/a(\boldsymbol{\theta}), \tag{2}$$

where $\boldsymbol{c}(\boldsymbol{\theta})$ is the canonical parameter vector and its coefficient is a (complete) sufficient statistic for $\boldsymbol{\theta}$.

(i) [2 marks]

Show that the score statistic (the gradient of the log likelihood for $\boldsymbol{\theta}$) can be expressed as

$$\partial \log L(\boldsymbol{\theta})/\partial \boldsymbol{\theta} \quad = \quad \boldsymbol{c}(\boldsymbol{\theta})^T [\boldsymbol{t}(\boldsymbol{x}) - E\{\boldsymbol{t}(\boldsymbol{X})\}], \tag{3}$$

where $\boldsymbol{X}$ denotes the random vector corresponding to $\boldsymbol{x}$.

(ii) [3 marks]

The ML estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is a solution of the equation

$$E\{\boldsymbol{t}(\boldsymbol{X})\} \quad = \quad \boldsymbol{t}(\boldsymbol{x}), \tag{4}$$

that is,

$$E_{\hat{\boldsymbol{\theta}}}\{\boldsymbol{t}(\boldsymbol{X})\} \quad = \quad \boldsymbol{t}(\boldsymbol{x}), \tag{5}$$

where $E_{\hat{\boldsymbol{\theta}}}$ denotes expectation using $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$.

Establish this result in the case where the parameter $\boldsymbol{\theta}$ is a scalar and is the canonical parameter; that is, $\boldsymbol{c}(\boldsymbol{\theta}) = \theta$.

(b) [3 marks]

Let $\boldsymbol{x}$ denote the complete-data vector in a EM framework adopted to calculate the ML estimate of a parameter $\boldsymbol{\theta}$ on the basis of some observed data $\boldsymbol{y}$ whose distribution is indexed by $\boldsymbol{\theta}$. If the complete-data likelihood $L_c(\boldsymbol{\theta})$ belongs to the regular exponential family, then on the M-step of the $(k+1)$th iteration the updated estimate $\boldsymbol{\theta}^{(k+1)}$ of $\boldsymbol{\theta}$ can be calculated as the solution of the equation

$$E\{\boldsymbol{t}(\boldsymbol{X})\} \quad = \quad \boldsymbol{t}^{(k)}(\boldsymbol{y}), \tag{6}$$

where

$$\boldsymbol{t}^{(k)}(\boldsymbol{y}) = E_{\boldsymbol{\theta}^{(k)}}\{\boldsymbol{t}(\boldsymbol{X}) \mid \boldsymbol{y}\}; \tag{7}$$

that is, $\boldsymbol{\theta}^{(k+1)}$ satisfies the equation

$$E_{\boldsymbol{\theta}^{(k+1)}}\{\boldsymbol{t}(\boldsymbol{X})\} \quad = \quad \boldsymbol{t}^{(k)}(\boldsymbol{y}). \tag{8}$$

Establish this result in the case where $\boldsymbol{c}(\boldsymbol{\theta}) = \theta$.

(c) [4 marks]

Suppose that a random sample of size n is taken from a two-component mixture density,

$$f(\boldsymbol{w}; \boldsymbol{\Psi}) = \sum_{i=1}^{2} \pi_i \, f_i(\boldsymbol{w}; \boldsymbol{\theta}_i), \tag{9}$$

where $\boldsymbol{\Psi} = (\pi_1, \boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$.

(i) [2 marks]

Show that the likelihood equation for $\boldsymbol{\Psi}$ can be manipulated into the form

$$\hat{\pi}_1 \quad = \quad \sum_{j=1}^{n} \tau_1(\boldsymbol{w}_j; \hat{\boldsymbol{\Psi}})/n, \tag{10}$$

$$\hat{\boldsymbol{\theta}}_i \quad = \quad \sum_{j=1}^{n} \tau_i(\boldsymbol{w}_j; \hat{\boldsymbol{\Psi}}) \, \partial \log f_i(\boldsymbol{w}_j; \hat{\boldsymbol{\theta}}_i)/\partial \boldsymbol{\theta}_i \quad (i = 1, 2), \tag{11}$$

where

$$\tau_i(\boldsymbol{w}_j; \hat{\boldsymbol{\Psi}}) = \frac{\hat{\pi}_i\, f_i(\boldsymbol{w}_j; \hat{\boldsymbol{\theta}}_i)}{f(\boldsymbol{w}_j; \hat{\boldsymbol{\Psi}})} \quad (i = 1, 2;\ j = 1, \ldots, n).$$

(ii) [2 marks]

Suggest a method for solving these equations iteratively to find the ML estimate $\hat{\boldsymbol{\Psi}}$ of $\boldsymbol{\Psi}$.

(d) [6 marks]

Let $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$ denote a random sample on the random variable $\boldsymbol{W}$, which has a g-component normal mixture density,

$$f(\boldsymbol{w}; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i\, \phi(\boldsymbol{w}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),$$

where $\phi(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the p-dimensional multivariate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Here

$$\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_g^T)^T,$$

where $\boldsymbol{\theta}_i$ contains the p elements of μ_i and the $\frac{1}{2}p(p+1)$ distinct elements of the covariance matrix $\boldsymbol{\Sigma}_i$ for $i = 1, \ldots, g$.

It is proposed to apply the EM algorithm to find the ML estimate of the parameter vector $\boldsymbol{\Psi}$ as a solution of the likelihood equation,

$$\partial \log L_(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = \mathbf{0},$$

where

$$\log L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \log \sum_{i=1}^{g} \pi_i\, \phi(\boldsymbol{w}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i).$$

In the EM framework adopted here, $\boldsymbol{w}_j$ is viewed as a realization of the random variable $\boldsymbol{W}_j$ drawn at random from a mixture of g Classes $C_1 \ldots, C_g$ in proportions $\pi_1, \ldots, \pi_g$ in which $\boldsymbol{W}$ has density $\phi(\boldsymbol{w}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ in Class C_i $(i = 1, \ldots, g)$. Accordingly, we introduce

$$\boldsymbol{z} = (\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_n^T)^T$$

as the missing-data vector, where $z_{ij} = (\boldsymbol{z}_j)_i$ is defined by

$$\begin{aligned} z_{ij} &= 1, \quad \text{if } \boldsymbol{w}_j \in C_i, \\ &= 0, \quad \text{otherwise,} \end{aligned}$$

for $i = 1, \ldots, g$; $j = 1, \ldots, n$.

Hence the complete-data vector $\boldsymbol{x}$ is given by

$$\boldsymbol{x} = (\boldsymbol{y}^T, \boldsymbol{z}^T)^T,$$

where $\boldsymbol{y} = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_n^T)^T$ denotes the (observed) incomplete-data vector.

(i) [2 marks]

Give the complete-data likelihood $L_c(\boldsymbol{\Psi})$ for $\boldsymbol{\Psi}$.

(ii) [2 marks]

Calculate the Q-function, $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$, on the E-step of the $(k+1)$th EM iteration.

(iii) [2 marks]

The form of the maximum likelihood (ML) estimate of $\boldsymbol{\Psi}$ obtained directly by differentiation of the complete-data log likelihood $\log L_c(\boldsymbol{\Psi})$ is known and it applies also when the known zero-one values of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are replaced by fractional values between zero and one.

Use this result to derive the updated estimate $\boldsymbol{\Psi}^{(k+1)}$ of $\boldsymbol{\Psi}$ on the M-step of the $(k+1)$th EM iteration.

(e) [12 marks]

Another approach to the problem in (d) for calculating the ML estimate of $\boldsymbol{\Psi}$ on the basis of the observed data $\boldsymbol{y}$ is to use the fact that the complete-data density $g_c(\boldsymbol{x}; \boldsymbol{\Psi})$ belongs to the regular exponential family. It can be shown that for $g_c(\boldsymbol{x}; \boldsymbol{\Psi}))$ in the regular exponential form (2), the sufficient statistic statistic $\boldsymbol{t}(\boldsymbol{X})$ contains in the present context of a g-component normal mixture model the $(g-1)$ elements,

$$\sum_{j=1}^{n} Z_{ij} \quad (i = 1, \ldots, g-1),$$

the p elements in each of the g vectors,

$$\sum_{j=1}^{n} Z_{ij}\, \boldsymbol{W}_j \quad (i = 1,\, \ldots,\, g),$$

and the $\frac{1}{2}p(p+1)$ distinct elements in each of the g symmetric matrices,

$$\sum_{j=1}^{n} Z_{ij}\, \boldsymbol{W}_j \boldsymbol{W}_j^T \quad (i = 1,\, \ldots,\, g).$$

(i) [4 marks]

Calculate $E(\boldsymbol{t}(\boldsymbol{X})\}$ and $E\{\boldsymbol{t}(\boldsymbol{X}) \mid \boldsymbol{y}\}$ from the above expressions for the elements of $\boldsymbol{t}(\boldsymbol{X})$.

Use the relationship (8) to derive the updated estimate $\boldsymbol{\Psi}^{(k+1)}$ of $\boldsymbol{\Psi}$ on the $(k+1)$th EM iteration.

(ii) [4 marks]

In the univariate case $(p = 1)$ of the two-component normal mixture density,

$$f(w; \boldsymbol{\Psi}) = \sum_{i=1}^{2} \pi_i\, \phi(w; \mu_i, \sigma_i^2),$$

where now

$$\boldsymbol{\Psi} = (\pi_1, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^T,$$

verify that the complete-data density $g_c(\boldsymbol{x}; \boldsymbol{\Psi})$ belongs to the regular exponential family.

Give explicitly the values of the canonical parameter $\boldsymbol{c}(\boldsymbol{\Psi})$, the sufficient statistic $\boldsymbol{t}(\boldsymbol{x})$, $a(\boldsymbol{\Psi})$, and $b(\boldsymbol{x})$.

(iii) [4 marks]

Verify the result that

$$E\{\boldsymbol{t}(\boldsymbol{X})\} = \partial \log a(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi},$$

using the expressions obtained above in e(ii).