UiO : Matematisk institutt
Det matematisk-naturvitenskapelige fakultet

STK-4051/9051 Computational Statistics Spring 2024
Ex4

Instructor: Odd Kolbjørnsen, oddkol@math.uio.no

Exercise 7 (Mixture of Gaussians)
Assume $\boldsymbol{Y}_i = (X_i, C_i)$ are distributed according to

$$\begin{aligned} \Pr(C_i = k) =& \pi_k, & k = 1, ..., K \\ X_i | C_i = k \sim& N(\mu_k, \sigma_k^2) \end{aligned}$$

but where the C_i's are missing. The complete log-density for a single observation $\boldsymbol{y}_i$ is given by

$$\begin{aligned} \log f(\boldsymbol{y}_i) =& \log(\pi_{c_i}) + \log[\phi(x_i; \mu_{c_i}, \sigma_{c_i}^2)] \\ =& \sum_{k=1}^{K} I(c_i = k)[\log(\pi_k) + \log[\phi(x_i; \mu_k, \sigma_k^2)]] \end{aligned}$$

while the complete log-likelihood:

$$\log f_Y(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} I(c_i = k)[\log(\pi_k) + \log[\phi(x_i; \mu_k, \sigma_k^2)]$$

The E-step (taking into account that the C_i's are the only stochastic parts) gives

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) =& E\{\sum_{i=1}^{n} \sum_{k=1}^{K} I(C_i = k)[\log(\pi_k) + \log[\phi(x_i; \mu_k, \sigma_k^2)]|\boldsymbol{x}, \boldsymbol{\theta}^{(t)}\} \\ =& \sum_{i=1}^{n} \sum_{k=1}^{K} \Pr(C_i = k|\boldsymbol{x}, \boldsymbol{\theta}^{(t)})[\log(\pi_k) + \log[\phi(x_i; \mu_k, \sigma_k^2)] \end{aligned}$$

Show that the M-step in the EM algorithm corresponds to

$$\begin{aligned} \pi_k^{(t+1)} =& \frac{1}{n} \sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{x}, \boldsymbol{\theta}^{(t)}) \\ \mu_k^{(t+1)} =& \frac{1}{n\pi_k^{(t+1)}} \sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{x}, \boldsymbol{\theta}^{(t)}) x_i \\ (\sigma_k^2)^{(t+1)} =& \frac{1}{n\pi_k^{(t+1)}} \sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{x}, \boldsymbol{\theta}^{(t)})(x_i - \mu_k^{(t+1)})^2 \end{aligned}$$

Solution to exercise 7.

Since $\sum_k \pi_k = 1$, we need to introduce a Lagrange term:

$$Q_{lagr}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \Pr(C_i = k|\boldsymbol{x}, \boldsymbol{\theta}^{(t)})[\log(\pi_k) - \tfrac{1}{2}\log(\sigma_k^2) - \tfrac{1}{2\sigma_k^2}(x_i - \mu_k)^2] + \lambda(1 - \sum_{k=1}^{K}\pi_k)$$

$$\frac{\partial}{\partial \pi_k} Q_{lagr}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{x}, \boldsymbol{\theta}^{(t)})\pi_k^{-1} - \lambda$$

giving

$$\begin{aligned}\pi_k^{(t+1)} &= \lambda^{-1}\sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{x}, \boldsymbol{\theta}^{(t)}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{x}, \boldsymbol{\theta}^{(t)})\end{aligned}$$

Further

$$\frac{\partial}{\partial \mu_k} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{x}, \boldsymbol{\theta}^{(t)})[\tfrac{1}{\sigma_k^2}(x_i - \mu_k)]$$

giving

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{x}, \boldsymbol{\theta}^{(t)})x_i}{\sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{x}, \boldsymbol{\theta}^{(t)})}$$

Similarly,

$$\frac{\partial}{\partial \sigma_k^2} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{x}, \boldsymbol{\theta}^{(t)})[-\tfrac{1}{2\sigma_k^2} + \tfrac{1}{2\sigma_k^4}(x_i - \mu_k)^2]$$

giving

$$(\sigma_k^2)^{(t+1)} = \frac{\sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{x}, \boldsymbol{\theta}^{(t)})(x_i - \mu_k^{(t+1)})^2}{\sum_{i=1}^{n} \Pr(C_i = k|\boldsymbol{x}, \boldsymbol{\theta}^{(t)})}$$

4.2. Epidemiologists are interested in studying the sexual behavior of individuals at risk for HIV infection. Suppose 1500 gay men were surveyed and each was asked how many risky sexual encounters he had in the previous 30 days. Let n_i denote the number of respondents reporting i encounters, for $i = 1, \ldots, 16$. Table 4.2 summarizes the responses.

These data are poorly fitted by a Poisson model. It is more realistic to assume that the respondents comprise three groups. First, there is a group of people who, for whatever reason, report zero risky encounters even if this is not true. Suppose a respondent has probability α of belonging to this group.

With probability β, a respondent belongs to a second group representing typical behavior. Such people respond truthfully, and their numbers of risky encounters are assumed to follow a Poisson(μ) distribution.

Finally, with probability $1 - \alpha - \beta$, a respondent belongs to a high-risk group. Such people respond truthfully, and their numbers of risky encounters are assumed to follow a Poisson(λ) distribution.

The parameters in the model are α, β, μ, and λ. At the tth iteration of EM, we use $\theta^{(t)} = (\alpha^{(t)}, \beta^{(t)}, \mu^{(t)}, \lambda^{(t)})$ to denote the current parameter values. The likelihood of the observed data is given by

$$L(\theta|n_0, \ldots, n_{16}) \propto \prod_{i=0}^{16} \left[\frac{\pi_i(\theta)}{i!}\right]^{n_i}, \tag{4.81}$$

where

$$\pi_i(\theta) = \alpha 1_{\{i=0\}} + \beta\mu^i \exp\{-\mu\} + (1 - \alpha - \beta)\lambda^i \exp\{-\lambda\} \tag{4.82}$$

$i = 0,1, \ldots, 16$

The observed data are $n_0, \ldots, n_{16}$. The complete data may be construed to be $n_{z,0}, n_{t,0}, \ldots, n_{t,16}$, and $n_{p,0}, \ldots, n_{p,16}$, where $n_{k,i}$ denotes the number of respondents in group k reporting i risky encounters and $k = z, t$, and p correspond to the zero, typical, and promiscuous groups, respectively. Thus, $n_0 = n_{z,0} + n_{t,0} + n_{p,0}$ and $n_i = n_{t,i} + n_{p,i}$ for $i = 1, \ldots, 16$. Let $N = \sum_{i=0}^{16} n_i = 1500$.

Define

$$z_0(\boldsymbol{\theta}) = \frac{\alpha}{\pi_0(\boldsymbol{\theta})}, \tag{4.83}$$

$$t_i(\boldsymbol{\theta}) = \frac{\beta\mu^i \exp\{-\mu\}}{\pi_i(\boldsymbol{\theta})}, \tag{4.84}$$

$$p_i(\boldsymbol{\theta}) = \frac{(1-\alpha-\beta)\lambda^i \exp\{-\lambda\}}{\pi_i(\boldsymbol{\theta})} \tag{4.85}$$

for $i = 0, \ldots, 16$. These correspond to probabilities that respondents with i risky encounters belong to the various groups.

$$\pi_i(\boldsymbol{\theta}) = \alpha 1_{\{i=0\}} + \beta\mu^i \exp\{-\mu\} + (1-\alpha-\beta)\lambda^i \exp\{-\lambda\}$$

a. Show that the EM algorithm provides the following updates:

$$\alpha^{(t+1)} = \frac{n_0 z_0(\theta^{(t)})}{N}, \tag{4.86}$$

$$\beta^{(t+1)} = \sum_{i=0}^{16} \frac{n_i t_i(\theta^{(t)})}{N}, \tag{4.87}$$

$$\mu^{(t+1)} = \frac{\sum_{i=0}^{16} i\, n_i t_i(\theta^{(t)})}{\sum_{i=0}^{16} n_i t_i(\theta^{(t)})}, \tag{4.88}$$

$$\lambda^{(t+1)} = \frac{\sum_{i=0}^{16} i\, n_i p_i(\theta^{(t)})}{\sum_{i=0}^{16} n_i p_i(\theta^{(t)})}. \tag{4.89}$$

The complete and marginal likelihood

Marginal likelihood

$$L(\boldsymbol{\theta}|n_0,\ldots,n_{16}) \propto \prod_{i=0}^{16}\left[\frac{\pi_i(\boldsymbol{\theta})}{i!}\right]^{n_i}, \qquad (4.81)$$

$$\pi_i(\boldsymbol{\theta}) = \alpha 1_{\{i=0\}} + \beta\mu^i \exp\{-\mu\} + (1-\alpha-\beta)\lambda^i \exp\{-\lambda\} \qquad (4.82)$$

The complete likelihood, we know also how many there are in each group (i.e. for each person we know group membership and count)

$$L\left(\theta \middle| n_{z,0}, n_{t,0}, \ldots, n_{t,16}, n_{p,0}, \ldots, n_{p,16}\right) \propto$$

$$\alpha^{n_{z,0}} \prod_{i=0}^{16}\left(\frac{\beta\mu^i \exp\{-\mu\}}{i!}\right)^{n_{t,i}}\left(\frac{\gamma\,\lambda^i \exp\{-\lambda\}}{i!}\right)^{n_{p,i}}$$

$$\gamma = 1-\alpha-\beta$$

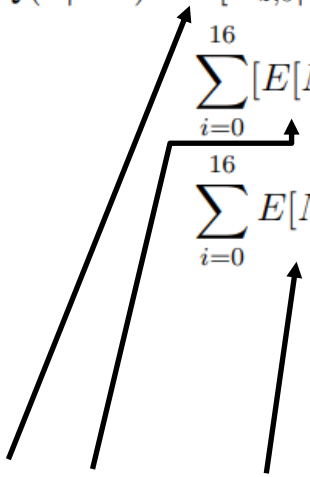
Solution to exercise (4.2)

(a). We introduce $\gamma = 1-\alpha-\beta$ with the constraints $\alpha+\beta+\gamma = 1$. Complete likelihood:

$$l(\boldsymbol{\theta}) = n_{z,0}\log(\alpha) + \sum_{i=0}^{16}[n_{t,i}(\log(\beta) + i\log(\mu) - \mu) + n_{p,i}(\log(\gamma) + i\log(\lambda) - \lambda)]$$

Then (with $\boldsymbol{n} = (n_0, ..., n_{16})$ and using s to denote iteration number in order not to confuse with t in model)

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = E[N_{z,0}|\boldsymbol{n}, \boldsymbol{\theta}^{(s)}]\log(\alpha) +$$
$$\sum_{i=0}^{16}[E[N_{t,i}|\boldsymbol{n}, \boldsymbol{\theta}^{(s)}](\log(\beta) + i\log(\mu) - \mu) +$$
$$\sum_{i=0}^{16} E[N_{p,i}|\boldsymbol{n}, \boldsymbol{\theta}^{(s)}](\log(\gamma) + i\log(\lambda) - \lambda)]$$



So we need the expectations

Assume now an individual j has answered i and denote by G_j the group membership. Then

$$\begin{aligned}
&\Pr(G_j = t | x_j = i) \\
\beta \rightarrow &= \frac{\Pr(G_i = t)\Pr(x_j = i | G_j = t)}{\Pr(x_j = i)} \\
&= \frac{\beta \mu^i \exp(-\mu)}{\pi_i(\boldsymbol{\theta})}
\end{aligned}$$

$\frac{\mu^i \exp(-\mu)}{i!}$

$\frac{\pi_i(\boldsymbol{\theta})}{i!}$

$G_j \in \{z, t, p\}$

($i!$ is then deleted in both the nominator and the denominator) which leads to (using that the individuals are independent so that $N_{t,i}$ is binomial distributed with probability defined above)

$$E[N_{t,i} | \boldsymbol{n}, \boldsymbol{\theta}^{(s)}] = n_i \frac{\beta^{(s)} (\mu^{(s)})^i \exp(-\mu^{(s)})}{\pi_i(\boldsymbol{\theta}^{(s)})} = n_i t_i(\boldsymbol{\theta}^{(s)})$$

We similarly get

$$E[N_{z,0} | \boldsymbol{n}, \boldsymbol{\theta}^{(s)}] = n_0 \frac{\alpha^{(s)}}{\pi_0(\boldsymbol{\theta}^{(s)})} = n_0 z_0(\boldsymbol{\theta}^{(s)})$$

$$E[N_{p,i} | \boldsymbol{n}, \boldsymbol{\theta}^{(s)}] = n_i \frac{\gamma^{(s)} (\lambda^{(s)})^i \exp(-\lambda^{(s)})}{\pi_i(\boldsymbol{\theta}^{(s)})} = n_i p_i(\boldsymbol{\theta}^{(s)})$$

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = E[N_{z,0}|\boldsymbol{n},\boldsymbol{\theta}^{(s)}]\log(\alpha) + \sum_{i=0}^{16}[E[N_{t,i}|\boldsymbol{n},\boldsymbol{\theta}^{(s)}](\log(\beta) + i\log(\mu) - \mu) + \sum_{i=0}^{16} E[N_{p,i}|\boldsymbol{n},\boldsymbol{\theta}^{(s)}](\log(\gamma) + i\log(\lambda) - \lambda)]$$

Lagrange multiplier

Further, introducing the Lagrange term,

$$Q_{lagr}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) + \phi(1 - \alpha - \beta - \gamma)$$

we get

$$\frac{\partial}{\partial\alpha}Q_{lagr}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = n_0 z_0(\boldsymbol{\theta}^{(s)})\frac{1}{\alpha} - \phi$$

so

$$\alpha^{(s+1)} = \frac{1}{\phi}n_0 z_0(\boldsymbol{\theta}^{(s)})$$

$$\frac{\partial}{\partial\beta}Q_{lagr}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(s)})\frac{1}{\beta} - \phi$$

so

$$\beta^{(s+1)} = \frac{1}{\phi}\sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(s)})$$

$$\frac{\partial}{\partial\gamma}Q_{lagr}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \sum_{i=0}^{16} p_i(\boldsymbol{\theta}^{(s)})\frac{1}{\gamma} - \phi$$

so

$$\gamma^{(s+1)} = \frac{1}{\phi}\sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^{(s)})$$

By noting that

$$n_0 z_0(\boldsymbol{\theta}^{(s)}) + \sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(s)}) + \sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^{(s)}) = N$$

we get: $\phi = N$

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = E[N_{z,0}|\boldsymbol{n}, \boldsymbol{\theta}^{(s)}] \log(\alpha) + \sum_{i=0}^{16} [E[N_{t,i}|\boldsymbol{n}, \boldsymbol{\theta}^{(s)}](\log(\beta) + i\log(\mu) - \mu) + \sum_{i=0}^{16} E[N_{p,i}|\boldsymbol{n}, \boldsymbol{\theta}^{(s)}](\log(\gamma) + i\log(\lambda) - \lambda)]$$

Further, introducing the Lagrange term,

$$Q_{lagr}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) + \phi(1 - \alpha - \beta - \gamma)$$

$$\frac{\partial}{\partial \mu} Q_{lagr}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(s)})(\frac{i}{\mu} - 1)$$

$$\mu^{(s+1)} = \frac{\sum_{i=0}^{16} i n_i t_i(\boldsymbol{\theta}^{(s)})}{\sum_{i=0}^{16} n_i t_i(\boldsymbol{\theta}^{(s)})}$$

and similarly

$$\lambda^{(s+1)} = \frac{\sum_{i=0}^{16} i n_i p_i(\boldsymbol{\theta}^{(s)})}{\sum_{i=0}^{16} n_i p_i(\boldsymbol{\theta}^{(s)})}$$

b. Estimate the parameters of the model, using the observed data.

```
alpha = 0.6
beta = 0.3
mu = 1
lambda = 10
i = 0:16
eps = 0.001
l = loglik(alpha,beta,mu,lambda,x)
more = TRUE
while(more)
{
 l.old = l
 pi = (beta*exp(-mu)*mu^i + (1-alpha-beta)*exp(-lambda)*lambda^i)
 pi[1] = pi[1]+alpha
 zstat0 = alpha/pi[1]
 tstat = beta*exp(-mu)*mu^i/pi
 pstat = (1-alpha-beta)*exp(-lambda)*lambda^i/pi
 alpha = x$freq[1]*zstat0/N
 beta = sum(x$freq*tstat)/N
 mu = sum(i*x$freq*tstat)/sum(x$freq*tstat)
 lambda = sum(i*x$freq*pstat)/sum(x$freq*pstat)
 param = c(log(alpha/(1-alpha)),log(beta/(1-beta)),log(mu),log(1-alpha-beta))
 l = loglik(alpha,beta,mu,lambda,x)
 more = abs(l-l.old)>eps
 show(c(alpha,beta,mu,lambda,l))
}
print("Estimates from the EM algorithm")
show(c(alpha,beta,mu,lambda))
[1] 0.1236868 0.5625885 1.4771350 5.9510163
```

What happends if you start off:
mu=5
lambda=5

Problem 2 (EM algorithm)

Consider the mixture model for clustering:

$$P(C_i = k) = \frac{1}{K}, \quad k = 1, \dots, K, i = 1, \dots, n$$

$$p(x_i | C_i = k) = \phi(x; \mu_k, \sigma_k^2), \qquad k = 1, .. K, i = 1, \dots, n$$

Where $x = (x_1, \dots x_n)$ is the observations, $C = (C_1, \dots C_n)$ is the class labels, $\phi(x; \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2. Our aim is to obtain maximum likelihood estimates of $\theta = \{\mu_k, \sigma_k^2, k = 1, \dots, K\}$ based on observations $x = (x_1, \dots x_n)$. The class labels are missing.

a) In the context of the EM algorithm write down the expression for the complete log-likelihood, derive the expression for $Q(\theta | \theta^{(t)})$, and show that the update on μ_k , and σ_k^2, is given by:

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^n P(C_i = k | x_i, \theta^{(t)}) x_i}{\sum_{i=1}^n P(C_i = k | x_i, \theta^{(t)})}$$

$$(\sigma_k^2)^{(t+1)} = \frac{\sum_{i=1}^n P(C_i = k | x_i, \theta^{(t)}) \left(x_i - \mu_k^{(t+1)}\right)^2}{\sum_{i=1}^n P(C_i = k | x_i, \theta^{(t)})}$$

Derive also the expression for $P(C_i = k | x_i, \theta^{(t)})$.

The complete log likelihood

$$l(\theta|\boldsymbol{x},\boldsymbol{c}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \log\frac{1}{K} I(c_i = k) \log \phi\left(x_i; \mu_k, \sigma_k^2\right)$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} \log\frac{1}{K} I(c_i = k) \left(-\frac{1}{2}\log\left(2\pi\sigma_k^2\right) - \frac{1}{2}\frac{(x_i - \mu_k)^2}{\sigma_k^2}\right)$$

$$Q\left(\theta\middle|\theta^{(t)}\right) = E\left(l(\theta|\boldsymbol{x},\boldsymbol{c})\middle|\boldsymbol{x},\theta^{(t)}\right)$$

$$= \text{Const} + \sum_{i=1}^{n}\sum_{k=1}^{K} P\left(C_i = k|\boldsymbol{x},\theta^{(t)}\right)\left(-\frac{1}{2}\log\left(\sigma_k^2\right) - \frac{1}{2}\frac{(x_i - \mu_k)^2}{\sigma_k^2}\right)$$

$$\frac{\partial Q\left(\theta\middle|\theta^{(t)}\right)}{\partial \mu_k} = \sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x},\theta^{(t)}\right)\left(\frac{(x_i - \mu_k)}{\sigma_k^2}\right) = 0$$

$$\mu_k \sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x},\theta^{(t)}\right) = \sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x},\theta^{(t)}\right) x_i$$

$$\mu_k = \frac{\sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x},\theta^{(t)}\right) x_i}{\sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x},\theta^{(t)}\right)}$$

$$Q\left(\theta\middle|\theta^{(t)}\right) = E\left(l(\theta|\boldsymbol{x},\boldsymbol{c})\middle|\boldsymbol{x},\theta^{(t)}\right)$$
$$= \text{Const} + \sum_{i=1}^{n}\sum_{k=1}^{K} P\left(C_i = k|\boldsymbol{x},\theta^{(t)}\right)\left(-\frac{1}{2}\log\left(\sigma_k^2\right) - \frac{1}{2}\frac{(x_i-\mu_k)^2}{\sigma_k^2}\right)$$

$$\frac{\partial Q\left(\theta\middle|\theta^{(t)}\right)}{\partial\sigma_k^2} = \sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x},\theta^{(t)}\right)\left(-\frac{1}{2\sigma_k^2} + \frac{1}{2}\frac{(x_i-\mu_k)^2}{(\sigma_k^2)^2}\right) = 0$$

$$\sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x},\theta^{(t)}\right) = \frac{1}{\sigma_k^2}\sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x},\theta^{(t)}\right)(x_i-\mu_k)^2$$

$$\sigma_k^2 = \frac{\sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x},\theta^{(t)}\right)(x_i-\mu_k)^2}{\sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x},\theta^{(t)}\right)}$$

$$P\left(C_i = k|\boldsymbol{x},\theta^{(t)}\right) = \frac{p\left(C_i = k, X_i = x_i|\,\theta^{(t)}\right)}{p\left(X_i = x_i\middle|\theta^{(t)}\right)} = \frac{\frac{1}{K}\phi\left(x_i;\mu_k,\sigma_k^2\right)}{\frac{1}{K}\sum_{m=1}^{K}\phi\left(x_i;\mu_m,\sigma_m^2\right)} = \frac{\phi\left(x_i;\mu_k,\sigma_k^2\right)}{\sum_{m=1}^{K}\phi\left(x_i;\mu_m,\sigma_m^2\right)}$$

b) In semi supervised learning it is possible to enhance learning by actively observing the class membership of some of the observations, thus we get the additional information that $C_i = c_i$ for $i = 1, \dots, m$ where $m < n$. Given this additional information how would you change the updating rule for μ_k and σ_k^2 above. You do not need to show the full derivation of the new update, but comment on how the new information changes $Q(\theta|\theta^{(t)})$. Relate this change to the to the definition of $Q(\theta|\theta^{(t)})$.

--

b) The change is that the probability weight in the EM estimator becomes one if the class match the information, zero otherwise. Thus, for data with information of class we replace $P(C_i = k|\boldsymbol{x}, \theta^{(t)})$ with $I(C_i = c_i)$.

In terms of the expression for Q, we have that the expectation is taken over the observed data. Thus, when the class is observed, we get perfect information about the class, which gives the indicator function in the sum.

c) To assess the uncertainty in the EM estimator it is possible to use a bootstrap procedure. In the setting of the semi supervised learning from 2b describe both a parametric and a nonparametric bootstrap for assessing the uncertainty in the EM estimator. Discuss strengths and weaknesses in these two different approaches when applied to the problem of semi supervised learning in 2b.

c) Nonparametric bootstrap: Generate new data sets by resampling with replacement the data records from the original set. Positive: The sample is data driven, no assumption of the distributions are used. Negative: The number of known label classes will vary between samples.

Parametric bootstrap: Generate new data by random samples from the estimated model, apply the same label selection strategy as in the main set. Positive: We recreate the mechanism for selecting labels. Negative: We are limited to the model we have fitted.

In the remainder of the problem we will assume that $K = 2$ and that $(\sigma_1, \sigma_2) = (1,2)$. Thus, the unknown parameters in this problem is $\theta = (\mu_1, \mu_2)$ with $\mu_1 < \mu_2$. The histogram from one such model is shown in figure 1.

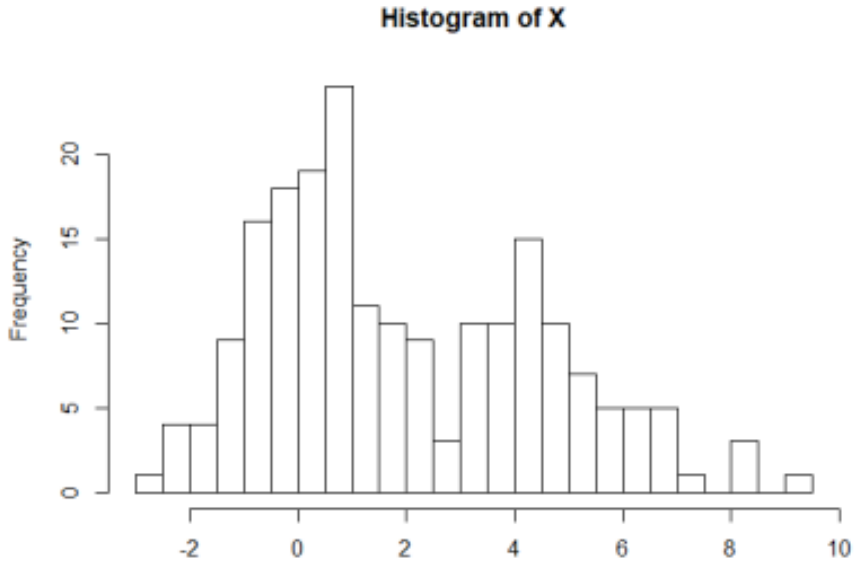


Figure 1: Histogram of the observations $x_1, \ldots, x_n$ used in problem 1d.

We will now consider the question about how to collect labels. Three different strategies are proposed:

A) Collect 10% of the labels at random
B) Collect the label from the 5% highest and 5% lowest values
C) Collect data from the 10% of the data closest to the median

A) Collect 10% of the labels at random
B) Collect the label from the 5% highest and 5% lowest values
C) Collect data from the 10% of the data closest to the median

d) Table 1 and 2 gives the values for the observed information matrix and its inverse, given the observations in figure 1. The cases shown are the unsupervised case from 1a, the three strategies A, B and C, and the complete likelihood where we know all class labels. Why is the inverse of the observed information matrix relevant? Based on the two tables, discuss which of the three models A, B and C that provide the most information, comment also on the result by comparing with the unsupervised and complete case. Which strategy for label selection would you use?

	Unsupervised		Strategy A		Strategy B		Strategy C		Complete	
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2
μ_1	74.26	-6.04	77.09	-5.34	76.54	-7.47	82.74	-2.55	99.00	0.00
μ_2	-6.04	16.29	-5.34	17.51	-7.47	17.04	-2.55	18.00	0.00	25.25

Table 1: Observed information matrix for the data in figure 1.

	Unsupervised		Strategy A		Strategy B		Strategy C		Complete	
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2
μ_1	0.014	0.005	0.013	0.004	0.014	0.006	0.012	0.002	0.010	0.000
μ_2	0.005	0.063	0.004	0.058	0.006	0.061	0.002	0.056	0.000	0.040

Table 2: The inverse of the observed information matrix for the data in figure 1.

	Unsupervised		Strategy A		Strategy B		Strategy C		Complete	
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2
μ_1	74.26	-6.04	77.09	-5.34	76.54	-7.47	82.74	-2.55	99.00	0.00
μ_2	-6.04	16.29	-5.34	17.51	-7.47	17.04	-2.55	18.00	0.00	25.25

Table 1: Observed information matrix for the data in figure 1.

	Unsupervised		Strategy A		Strategy B		Strategy C		Complete	
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2
μ_1	0.014	0.005	0.013	0.004	0.014	0.006	0.012	0.002	0.010	0.000
μ_2	0.005	0.063	0.004	0.058	0.006	0.061	0.002	0.056	0.000	0.040

Table 2: The inverse of the observed information matrix for the data in figure 1.

d) The inverse of the information matrix is an estimator of the sample covariance. Thus it is desirable to have large entries in the information matrix and small values in the inverse.

Strategy A gets some information, it is better than the strategy B, but still close to the unsupervised approach

Strategy B is the worst choice. When we select data from the edges this is where we already are quite certain about the classes, thus this brings little information compared with the unsupervised.

Strategy C gets the most information. This is natural since this gets labels from the region with large overlap in the distribution. We get information closer to the complete information.

The information matrices are sorted in terms of information content

$$\text{Unsupervised} < \text{Strategy B} < \text{Strategy A} < \text{Strategy C} < \text{Complete}$$

Exercise 9 (Preliminaries for the stochastic gradient algorithm)
We will in this exercise shows some preliminary results that will be used in deriving properties of the stochastic gradient algorithm.

(a). Assume $\{a_t\}$ is a series of finite and non-negative numbers such that

$$\sum_{t=1}^{\infty} a_t = \infty.$$

Show that $\sum_{t=T}^{\infty} a_t = \infty$ for any $T \geq 1$.

(a). We have

$$\infty = \sum_{t=1}^{\infty} a_t = \sum_{t=1}^{T} a_t + \sum_{t=T+1}^{\infty} a_t$$

Since the first sum on the right hand side is finite, the second has to be infinite.

(b). Assume $\alpha_t > 0$ and

$$\sum_{t=2}^{\infty} \frac{\alpha_t}{\alpha_1 + \cdots + \alpha_{t-1}} = \infty$$

Show that then $\sum_{t=1}^{\infty} \alpha_t = \infty$.

(b). Assume $\sum_{t=1}^{\infty} \alpha_t < \infty$. Then

$$\sum_{t=2}^{\infty} \frac{\alpha_t}{\alpha_1 + \cdots + \alpha_{t-1}} \leq \sum_{t=2}^{\infty} \frac{\alpha_t}{\alpha_1} = \frac{1}{\alpha_1} \sum_{t=2}^{\infty} \alpha_t < \infty.$$

giving a contradiction.

(c). Assume $\{a_t\}$ and $\{b_t\}$ are two series of finite and non-negative numbers such that $\lim_{t\to\infty} b_t$ exists and

$$\sum_{t=1}^{\infty} a_t = \infty, \quad \sum_{t=1}^{\infty} a_t b_t < \infty$$

Show that then $\lim_{t\to\infty} b_t = 0$.

(c). Assume $\lim_{t\to\infty} b_t = \delta > 0$. Then there exists some $0 < \varepsilon < \delta$ and T such that $b_t > \varepsilon$ for $t > T$. Then

$$\sum_{t=1}^{\infty} a_t b_t = \sum_{t=1}^{T} a_t b_t + \sum_{t=T+1}^{\infty} a_t b_t \geq \sum_{t=1}^{T} a_t b_t + \varepsilon \sum_{t=T+1}^{\infty} a_t = \infty$$

giving a contradiction.

(d). Consider a sequence

$$\theta^{t+1} = \theta^t - \alpha_t Z(\theta^t, \xi^t)$$

where $|Z(\theta^t, \xi^t)| < C$ with probability one. Show that

$$|\theta^t - \theta^*| \leq A_t \equiv |\theta^1 - \theta^*| + C(\alpha_1 + \cdots + \alpha_{t-1}).$$

(d). We have that

$$\begin{aligned}
\theta^t =& \theta^{t-1} - \alpha_{t-1} Z(\theta^{t-1}, \xi^{t-1}) \\
=& \theta^{t-2} - \alpha_{t-2} Z(\theta^{t-2}, \xi^{t-2}) - \alpha_{t-1} Z(\theta^{t-1}, \xi^{t-1}) \\
& \vdots \\
=& \theta^1 - \sum_{s=1}^{t-1} \alpha_s Z(\theta^s, \xi^s)
\end{aligned}$$

giving

$$\begin{aligned}
|\theta^t - \theta^*| =& |\theta^1 - \sum_{s=1}^{t-1} \alpha_s Z(\theta^s, \xi^s) - \theta^*| \leq |\theta^1 - \theta^*| + \sum_{s=1}^{t-1} \alpha_s |Z(\theta^s, \xi^s)| \\
\leq& |\theta^1 - \theta^*| + \sum_{s=1}^{t-1} \alpha_s C = A_t
\end{aligned}$$