B555 - Machine Learning - Homework 2

Enrique Areyan
March 05, 2015

Problem 1: Suppose that the number of accidents occurring daily in a certain plant has a Poisson distribution with an unknown mean λ. Based on previous experience in similar industrial plants, suppose that our initial feeling about the possible value of λ can be expressed by an exponential distribution with parameter $\theta = \frac{1}{2}$. That is, the prior density is

$$f(\lambda) = \theta e^{-\theta\lambda}$$

where $\lambda \in (0, \infty)$. If there are 79 accidents over the next 9 days, determine:

a) the maximum likelihood estimate of λ.

b) the maximum a posteriori estimate of λ.

c) the Bayes estimate of λ

Solution: This situation can be modeled as having data set $\mathcal{D} = \{x_i\}_{i=1}^{9}$, where each x_i = number of accidents in the plant on day i, for $1 \leq i \leq 9$. We do not have the value of each x_i, but we know that $\sum_{i=1}^{9} x_i = 79$.

a) Maximum Likelihood: by definition:

$$\lambda_{ML} = \arg\max_{\lambda}\{p(\mathcal{D}|\lambda)\}$$

where the probability of a single observation x_i given λ is $p(x_i|\lambda) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$, since we assume a poisson distribution for the number of accidents. From this it follows that the likelihood for a general the data set $\mathcal{D}$ with n observations is:

$$\begin{aligned}
p(\mathcal{D}|\lambda) &= \prod_{i=1}^{n} p(x_i|\lambda) && \text{by independence of } x_i \\
&= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} && \text{by assumption of Poisson distribution} \\
&= \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} && \text{arithmetic}
\end{aligned}$$

This shows that the likelihood function is $l(\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$. Instead of maximizing this function, let us maximize the log-likelihood:

$$ll(\lambda) = log\left(\frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}\right) = log\left(\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}\right) - log\left(\prod_{i=1}^{n} x_i!\right) = log(\lambda)\sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} log(x_i!)$$

Maximize:

$$\frac{\partial ll}{\partial \lambda} = \frac{\partial}{\partial \lambda}\left[log(\lambda)\sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} log(x_i!)\right] = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n$$

Setting $\frac{\partial ll}{\partial \lambda} = 0$ we obtain $\lambda = \frac{\sum_{i=1}^{n} x_i}{n}$.

We can check that indeed this is a global maximum by checking the second derivative:

$$\frac{\partial^2 ll}{\partial \lambda^2} = \frac{\partial}{\partial \lambda}\left[\frac{\sum_{i=1}^{n} x_i}{\lambda} - n\right] = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2} < 0$$

Since $x_i \in \mathbb{N}$ for all i and $\lambda^2 > 0$. Here we ignore the degenerate case where $x_i = 0$ for all i.
Having check that we indeed have the global maximum, we can conclude:

$$\boxed{\lambda_{ML} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad \text{where } x_i \text{ are i.i.d observations from a Poisson distribution}}$$

In our case: $n = 9$ and $\sum_{i=1}^{n} x_i = 79$, hence $\boxed{\lambda_{ML} = \frac{79}{9}}$, so an average rate of $8\frac{7}{9}$.

b) Maximum a posteriori: by definition:

$$\lambda_{MAP} = \arg\max_{\lambda}\{p(\mathcal{D}|\lambda)p(\lambda)\}$$

where $p(\mathcal{D}|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$ as computed in part a, and $p(\lambda) = \theta e^{-\theta\lambda}$ by assumption. Thus,

$$p(\mathcal{D}|\lambda)p(\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \cdot \theta e^{-\theta\lambda} = \frac{\lambda^{\sum_{i=1}^{n} x_i} \theta e^{-\lambda(n+\theta)}}{\prod_{i=1}^{n} x_i!}$$

This function is to be maximized to obtain the maximum a posteriori. However, let us instead maximize the log of this function:

$$\begin{aligned}
log(p(\mathcal{D}|\lambda)p(\lambda)) &= log\left[\frac{\lambda^{\sum_{i=1}^{n} x_i} \theta e^{-\lambda(n+\theta)}}{\prod_{i=1}^{n} x_i!}\right] \\
&= log\left[\lambda^{\sum_{i=1}^{n} x_i} \theta e^{-\lambda(n+\theta)}\right] - log\left[\prod_{i=1}^{n} x_i!\right] \\
&= log(\lambda)\sum_{i=1}^{n} x_i + log(\theta) - \lambda(n+\theta) - \sum_{i=1}^{n} log(x_i!)
\end{aligned}$$

Maximize:

$$\frac{\partial}{\partial \lambda}\left[log(p(\mathcal{D}|\lambda)p(\lambda))\right] = \frac{\partial}{\partial \lambda}\left[log(\lambda)\sum_{i=1}^{n} x_i + log(\theta) - \lambda(n+\theta) - \sum_{i=1}^{n} log(x_i!)\right] = \frac{\sum_{i=1}^{n} x_i}{\lambda} - (n+\theta)$$

Setting $\frac{\partial}{\partial \lambda}\left[log(p(\mathcal{D}|\lambda)p(\lambda))\right] = 0$ we obtain $\lambda = \frac{\sum_{i=1}^{n} x_i}{n+\theta}$.

We can check that indeed this is a global maximum by checking the second derivative:

$$\frac{\partial^2}{\partial \lambda^2}\left[log(p(\mathcal{D}|\lambda)p(\lambda))\right] = \frac{\partial}{\partial \lambda}\left[\frac{\sum_{i=1}^{n} x_i}{\lambda} - (n+\theta)\right] = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2} < 0$$

Since $x_i \in \mathbb{N}$ for all i and $\lambda^2 > 0$. Here we ignore the degenerate case where $x_i = 0$ for all i. Having check that we indeed have the global maximum, we can conclude:

$$\boxed{\lambda_{MAP} = \frac{\sum\limits_{i=1}^{n} x_i}{n+\theta}, \qquad \text{where } x_i \text{ are i.i.d observations from a Poisson distribution and } \lambda \text{ has an exponential prior}}$$

In our case: $n = 9$, $\sum\limits_{i=1}^{n} x_i = 79$, and $\theta = 1/2$ hence $\boxed{\lambda_{MAP} = \frac{79}{9+1/2} = \frac{158}{19}}$, so an average rate of $8\frac{6}{19}$.

c) Bayes Estimate: by definition:

$$\lambda_B = E[\lambda|\mathcal{D}] = \int_0^\infty p(\lambda|\mathcal{D})\lambda d\lambda, \qquad \text{i.e., the mean of the posterior distribution}$$

where $p(\lambda|\mathcal{D}) = \frac{p(\mathcal{D}|\lambda)p(\lambda)}{p(\mathcal{D})}$. Let us compute each piece separately:

$$p(\mathcal{D}|\lambda)p(\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i}\theta e^{-\lambda(n+\theta)}}{\prod\limits_{i=1}^{n} x_i!}, \text{ wich we computed before}$$

To compute $p(D)$, we can marginalize over all values of λ:

$$\begin{aligned} p(D) &= \int_0^\infty p(\mathcal{D}|\lambda)p(\lambda)d\lambda && \text{marginalization} \\ &= \int_0^\infty \frac{\lambda^{\sum_{i=1}^{n} x_i}\theta e^{-\lambda(n+\theta)}}{\prod\limits_{i=1}^{n} x_i!} d\lambda && \text{computed before} \\ &= \int_0^\infty \frac{\lambda^{\sum_{i=1}^{n} x_i}\theta e^{-\lambda(n+\theta)}}{\prod\limits_{i=1}^{n} x_i!} d\lambda && \text{computed before} \end{aligned}$$

This integral is computable but there is an easier way. Instead of doing this integral, let us find the functional form of the posterior by a proportionality argument:

$$p(\mathcal{D}|\lambda)p(\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i}\theta e^{-\lambda(n+\theta)}}{\prod\limits_{i=1}^{n} x_i!} \propto \lambda^{\sum_{i=1}^{n} x_i} e^{-\lambda(n+\theta)} \quad , \text{ dropping all values that do not depend on } \lambda$$

This shows that the posterior follows a Gamma distribution. Recall (or see reference [1]), that if X has a Gamma distribution with parameters $\alpha, \beta \in (0, \infty)$ then X has a pdf proportional to $x^{\alpha-1}e^{-\beta x}$ and its mean is $E[X] = \frac{\alpha}{\beta}$.

Therefore, the posterior $p(\mathcal{D}|\lambda)p(\lambda)$ has a Gamma distribution with parameters $\alpha = \sum\limits_{i=1}^{n} x_i + 1$ and $\beta = n + \theta$. Now using the fact that we know what the mean of a Gamma distribution is, we can conclude:

$$\boxed{\lambda_B = \frac{\alpha}{\beta} = \frac{\sum\limits_{i=1}^{n} x_i + 1}{n+\theta}}$$

In our case: $n = 9$, $\sum\limits_{i=1}^{n} x_i = 79$, and $\theta = 1/2$ hence $\boxed{\lambda_B = \frac{79+1}{9+1/2} = \frac{160}{19}}$, so an average rate of $8\frac{8}{19}$.

Problem 2: Let $X_1, \ldots, X_n$ be i.i.d. Gaussian random variables, each having an unknown mean θ and known variance σ_0^2. If θ is itself selected from a normal population having a known mean μ and a known variance σ^2

a) what is the maximum a posteriori estimate of θ?

b) what is the Bayes estimate of θ?

Solution: a) Maximum a posteriori: by definition:

$$\theta_{MAP} = \arg\max_\theta \{p(\mathcal{D}|\theta)p(\theta)\}$$

where the probability of a single observation x_i given θ and σ_0^2 is $p(x_i|\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x_i-\theta)^2}{2\sigma_0^2}}$, since we assume a normal distribution for X_i. From this it follows that the likelihood for a general the data set $\mathcal{D}$ with n observations is:

$$\begin{aligned}
p(\mathcal{D}|\theta) &= \prod_{i=1}^n p(x_i|\theta) && \text{by independence of } x_i \\
&= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x_i-\theta)^2}{2\sigma_0^2}} && \text{by assumption of Normal distribution} \\
&= \frac{1}{(2\pi)^{n/2}\sigma_0^n} e^{-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma_0^2}} && \text{arithmetic}
\end{aligned}$$

Also, $p(\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}$.

Therefore,

$$\begin{aligned}
\theta_{MAP} &= \arg\max_\theta \{p(\mathcal{D}|\theta)p(\theta)\} \\
&= \arg\max_\theta \left\{ \frac{1}{(2\pi)^{n/2}\sigma_0^n} e^{-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma_0^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\theta-\mu)^2}{2\sigma^2}} \right\} \\
&= \arg\max_\theta \left\{ \frac{1}{(2\pi)^{(n+1)/2}\sigma_0^n\sigma} e^{-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma_0^2} - \frac{(\theta-\mu)^2}{2\sigma^2}} \right\}
\end{aligned}$$

As usual, let us take the log:

$$log(p(\mathcal{D}|\theta)p(\theta)) = log\left(\frac{1}{(2\pi)^{(n+1)/2}\sigma_0^n\sigma}\right) - \frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma_0^2} - \frac{(\theta-\mu)^2}{2\sigma^2}$$

Maximize:

$$\frac{\partial}{\partial\theta}\left[log(p(\mathcal{D}|\theta)p(\theta))\right] = \frac{\partial}{\partial\theta}\left[log\left(\frac{1}{(2\pi)^{(n+1)/2}\sigma_0^n\sigma}\right) - \frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma_0^2} - \frac{(\theta-\mu)^2}{2\sigma^2}\right] = \frac{\sum_{i=1}^n (x_i-\theta)}{\sigma_0^2} - \frac{(\theta-\mu)}{\sigma^2}$$

Setting $\frac{\partial}{\partial\theta}\left[log(p(\mathcal{D}|\theta)p(\theta))\right] = 0$ we obtain:

$$\begin{aligned}
0 &= \frac{\sum\limits_{i=1}^{n}(x_i - \theta)}{\sigma_0^2} - \frac{(\theta - \mu)}{\sigma^2} \\
&= \frac{(\sum\limits_{i=1}^{n} x_i) - n\theta}{\sigma_0^2} - \frac{\theta - \mu}{\sigma^2} \\
&= \frac{\sigma^2(\sum\limits_{i=1}^{n} x_i) - \sigma^2 n\theta - \sigma_0^2\theta + \sigma_0^2\mu}{\sigma_0^2\sigma^2} \\
&\implies \text{by canceling } \sigma_0^2\sigma^2 > 0 \\
0 &= \sigma^2(\sum\limits_{i=1}^{n} x_i) - \sigma^2 n\theta - \sigma_0^2\theta + \sigma_0^2\mu \\
&= -\theta(\sigma^2 n + \sigma_0^2) + \sigma^2(\sum\limits_{i=1}^{n} x_i) + \sigma_0^2\mu \\
&\implies \\
\theta &= \frac{\sigma^2(\sum\limits_{i=1}^{n} x_i) + \sigma_0^2\mu}{\sigma^2 n + \sigma_0^2}
\end{aligned}$$

We can check that indeed this is a global maximum by checking the second derivative:

$$\frac{\partial^2}{\partial\lambda^2}\left[log(p(\mathcal{D}|\theta)p(\theta))\right] = \frac{\partial}{\partial\theta}\left[\frac{(\sum\limits_{i=1}^{n} x_i) - n\theta}{\sigma_0^2} - \frac{\theta - \mu}{\sigma^2}\right] = -\frac{n}{\sigma_0^2} - \frac{1}{\sigma^2} < 0$$

Since $n > 0$ and $\sigma_0^2, \sigma^2 > 0$.
Having check that we indeed have the global maximum, we can conclude:

$$\boxed{\theta_{MAP} = \frac{\sigma^2(\sum\limits_{i=1}^{n} x_i) + \sigma_0^2\mu}{\sigma^2 n + \sigma_0^2} \qquad \text{where } x_i \text{ are i.i.d observations from a Normal distribution and } \theta \text{ has a Normal prior}}$$

Note that an equivalent way of writing this, which is found more commonly in the literature (See [2]) is:

$$\theta_{MAP} = \frac{\frac{\mu}{\sigma^2} + \frac{\sum\limits_{i=1}^{n} x_i}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{n}{\sigma_0^2}}$$

b) <u>Bayes Estimate</u>: by definition:

$$\theta_B = E[\theta|\mathcal{D}] = \int\limits_{-\infty}^{\infty} p(\theta|\mathcal{D})\theta d\theta, \qquad \text{i.e., the mean of the posterior distribution}$$

If we try to compute all the integrals we will most likely have a hard time solving them explicitly. Instead, as done in problem 1 part c), let us find the functional form of the posterior by a proportionality argument, i.e., by dropping all terms that do not depend on θ from $p(\mathcal{D}|\theta)p(\theta)$ we get:

$$p(\mathcal{D}|\theta)p(\theta) \propto exp\left\{\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{\sigma_0^2} + \frac{(\theta-\mu)^2}{\sigma^2}\right\}$$

This form is a Normal distribution but to show this clearly we will need to complete squares for θ ([3]):

$$\begin{aligned}
\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{\sigma_0^2} + \frac{(\theta-\mu)^2}{\sigma^2} &= \frac{\sum_{i=1}^{n}(x_i^2-2x_i\theta)+n\theta^2}{\sigma_0^2} + \frac{\theta^2-2\theta\mu+\mu^2}{\sigma^2} \\
&= \theta^2\left(\frac{n}{\sigma_0^2}+\frac{1}{\sigma^2}\right) - 2\theta\left(\frac{\mu}{\sigma^2}+\frac{\sum_{i=1}^{n}x_i}{\sigma_0^2}\right) + \left(\frac{\sum_{i=1}^{n}x_i^2}{\sigma_0^2}+\frac{\mu^2}{\sigma^2}\right) \\
&= \frac{1}{\sigma_1^2}(\theta-\mu_1)^2 + C
\end{aligned}$$

where C is a constant that does not depend on θ and $\mu_1 = \frac{\mu}{\sigma^2} + \frac{\sum_{i=1}^{n}x_i}{\sigma_0^2}$ and $\frac{1}{\sigma_1^2} = \frac{n}{\sigma_0^2} + \frac{1}{\sigma^2}$. Thus,

$$p(\mathcal{D}|\theta)p(\theta) \propto exp\{-\frac{1}{2\sigma_1^2}(\theta-\mu_1)^2\}$$

which means that the posterior distribution is normal with mean μ_1 and variance σ_1^2. The bayes estimate is just the mean of this distribution, i.e., μ_1:

$$\boxed{\theta_B = E[\theta|\mathcal{D}] = \mu_1 = \frac{\mu}{\sigma^2} + \frac{\sum_{i=1}^{n}x_i}{\sigma_0^2}}$$

Problem 3: Let $X_1, \ldots, X_n$ be i.i.d. random variables with distribution

$$f(x|\alpha) = \alpha^x(1-\alpha)^{1-x}$$

where $x \in (0,1)$. Assuming that the unknown parameter α was selected from a $(0,1)$ uniform distribution find the Bayes estimator of α.

Solution: Bayes Estimate: by definition:

$$\alpha_B = E[\alpha|\mathcal{D}] = \int_0^1 p(\alpha|\mathcal{D})\alpha d\theta, \qquad \text{i.e., the mean of the posterior distribution}$$

Let us find the functional form of the posterior distribution by a proportionality argument, i.e., by dropping all terms that do not depend on α from $p(\mathcal{D}|\alpha)p(\alpha)$ we get:

$$p(\mathcal{D}|\alpha)p(\alpha) \propto \left(\prod_{i=0}^{n}\alpha^{x_i}(1-\alpha)^{1-x_i}\right)\cdot 1 = \alpha^{\sum_{i=1}^{n}x_i}(1-\alpha)^{n-\sum_{i=1}^{n}x_i}$$

By looking at reference [4.], we see that the posterior belongs to the Beta distribution with parameters $\alpha = (\sum_{i=1}^{n}x_i)+1$ and $\beta = (n-\sum_{i=1}^{n}x_i)+1$. Since the mean of a Beta distribution is given by $\frac{\alpha}{\alpha+\beta}$, we conclude that:

$$\alpha_B = E[\alpha|\mathcal{D}] = \frac{\alpha}{\alpha+\beta} = \frac{(\sum_{i=1}^{n}x_i)+1}{(\sum_{i=1}^{n}x_i)+1+(n-\sum_{i=1}^{n}x_i)+1} = \boxed{\frac{(\sum_{i=1}^{n}x_i)+1}{n+2}}$$

Note that since $0 < \alpha < \alpha+\beta$ and $0 < \beta$, we have $0 < \alpha_B < 1$ as we are suppose to have.

Problem 4: Consider the following minimization problem:

$$\arg\min_{\mathbf{x}} ||\mathbf{A}\mathbf{x} - \mathbf{b}||$$

where $\mathbf{A}$ is a m-by-n, $\mathbf{x}$ is a n-by-1 vector and $\mathbf{b}$ is a m-by-1 vector (all vectors and matrices are real). Owing to the fact that the row space and nullspace of $\mathbf{A}$ are orthogonal, any vector $\mathbf{x} \in \mathbb{R}^n$ can be decomposed as $\mathbf{x} = \mathbf{x_r} + \mathbf{x_n}$, where $\mathbf{x_r}$ lies in the row space of $\mathbf{A}$ and $\mathbf{x_n}$ lies in the nullspace of $\mathbf{A}$. Suppose now that $\hat{\mathbf{x}} = \hat{\mathbf{x}}_\mathbf{r} + \hat{\mathbf{x}}_\mathbf{n}$ is one solution to the minimization problem above.

a) Prove that $\hat{\mathbf{x}} = \hat{\mathbf{x}}_\mathbf{r} + \alpha\hat{\mathbf{x}}_\mathbf{n}$, where $\alpha \in \mathbb{R}$, is also a solution to the minimization problem.

b) Prove that $\hat{\mathbf{x}}_\mathbf{r}$ from above is common to all solutions that minimize $||\mathbf{A}\mathbf{x}-\mathbf{b}||$. In other words, prove that there is no other vector from the row space that can be combined with any vector from the nullspace to minimize $||\mathbf{A}\mathbf{x} - \mathbf{b}||$

Solution: a) By hypothesis, $\hat{\mathbf{x}} = \hat{\mathbf{x}}_\mathbf{r} + \hat{\mathbf{x}}_\mathbf{n}$ is one solution to the minimization problem above. Hence, we have that the optimal value $d \in \mathbb{R}$ can be written as:

$$\begin{aligned}
d &= ||\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}|| && \text{since } d \text{ is the optimal}\\
&= ||\mathbf{A}(\hat{\mathbf{x}}_\mathbf{r} + \hat{\mathbf{x}}_\mathbf{n}) - \mathbf{b}|| && \text{by definition of } \hat{\mathbf{x}}\\
&= ||\mathbf{A}\hat{\mathbf{x}}_\mathbf{r} + \mathbf{A}\hat{\mathbf{x}}_\mathbf{n} - \mathbf{b}|| && \text{since } \mathbf{A} \text{ is a linear transformation (a matrix)}\\
&= ||\mathbf{A}\hat{\mathbf{x}}_\mathbf{r} + \mathbf{0} - \mathbf{b}|| && \text{since } \hat{\mathbf{x}}_\mathbf{n} \text{ is in the nullspace of } \mathbf{A}\\
&= ||\mathbf{A}\hat{\mathbf{x}}_\mathbf{r} - \mathbf{b}|| && \text{matrix and vector arithmetic}
\end{aligned}$$

This shows that the optimal value d can be written as $d = ||\mathbf{A}\hat{\mathbf{x}}_\mathbf{r} - \mathbf{b}||$. But consider:

$$\begin{aligned}
||\mathbf{A}(\hat{\mathbf{x}}_\mathbf{r} + \alpha\hat{\mathbf{x}}_\mathbf{n}) - \mathbf{b}|| &= ||\mathbf{A}\hat{\mathbf{x}}_\mathbf{r} + \alpha\mathbf{A}\hat{\mathbf{x}}_\mathbf{n} - \mathbf{b}|| && \text{since } \mathbf{A} \text{ is a linear transformation (a matrix)}\\
&= ||\mathbf{A}\hat{\mathbf{x}}_\mathbf{r} + \alpha\mathbf{0} - \mathbf{b}|| && \text{since } \hat{\mathbf{x}}_\mathbf{n} \text{ is in the nullspace of } \mathbf{A}\\
&= ||\mathbf{A}\hat{\mathbf{x}}_\mathbf{r} - \mathbf{b}|| && \text{matrix and vector arithmetic}\\
&= d && \text{by previous argument}
\end{aligned}$$

Therefore, $\hat{\mathbf{x}} = \hat{\mathbf{x}}_\mathbf{r} + \alpha\hat{\mathbf{x}}_\mathbf{n}$ is also a solution since it yields the optimal value d.

b) Let $\mathbf{p}$ be the projection of $\mathbf{b}$ to $C(\mathbf{A})$. Geometrically we know that this is a solution to $\arg\min_{\mathbf{x}} ||\mathbf{A}\mathbf{x}-\mathbf{b}||$. In other words, the projection $\mathbf{p}$ is the closest vector to $\mathbf{b}$ formed by vectors from $C(\mathbf{A})$. Since by hypothesis $\hat{\mathbf{x}} = \hat{\mathbf{x}}_\mathbf{r} + \hat{\mathbf{x}}_\mathbf{n}$ in one solution, we must have:

$$\mathbf{p} = \mathbf{A}(\hat{\mathbf{x}}_\mathbf{r} + \hat{\mathbf{x}}_\mathbf{n}) = \mathbf{A}\hat{\mathbf{x}}_\mathbf{r} + \mathbf{A}\hat{\mathbf{x}}_\mathbf{n} = \mathbf{A}\hat{\mathbf{x}}_\mathbf{r} + \mathbf{0} = \mathbf{A}\hat{\mathbf{x}}_\mathbf{r} \Longrightarrow \mathbf{p} = \mathbf{A}\hat{\mathbf{x}}_\mathbf{r} \qquad (*)$$

Suppose now that there is another vector $\mathbf{x}_\mathbf{r}^*$ from the row space, where $\mathbf{x}_\mathbf{r}^* \neq \hat{\mathbf{x}}_\mathbf{r}$, that can be combined with any vector from the nullspace $\mathbf{x}_\mathbf{n}^*$ to minimize $||\mathbf{A}\mathbf{x} - \mathbf{b}||$. In symbols, we have that:

$$\mathbf{p} = \mathbf{A}(\mathbf{x}_\mathbf{r}^* + \mathbf{x}_\mathbf{n}^*) = \mathbf{A}\mathbf{x}_\mathbf{r}^* + \mathbf{A}\mathbf{x}_\mathbf{n}^* = \mathbf{A}\mathbf{x}_\mathbf{r}^* + \mathbf{0} = \mathbf{A}\mathbf{x}_\mathbf{r}^* \Longrightarrow \mathbf{p} = \mathbf{A}\mathbf{x}_\mathbf{r}^* \qquad (**)$$

Subtracting equation $(*)$ from $(**)$:

$$\mathbf{p} - \mathbf{p} = \mathbf{A}\mathbf{x}_\mathbf{r}^* - \mathbf{A}\hat{\mathbf{x}}_\mathbf{r} \Longrightarrow \mathbf{0} = \mathbf{A}(\mathbf{x}_\mathbf{r}^* - \hat{\mathbf{x}}_\mathbf{r})$$

This means that we have found a vector $\mathbf{x}_\mathbf{r}^* - \hat{\mathbf{x}}_\mathbf{r} \neq 0$ (since $\mathbf{x}_\mathbf{r}^* \neq \hat{\mathbf{x}}_\mathbf{r}$) that belongs to the nullspace of $\mathbf{A}$. However, by hypothesis, both $\mathbf{x}_\mathbf{r}^*$ and $\hat{\mathbf{x}}_\mathbf{r}$ belong to the rowspace of $\mathbf{A}$. We know that any linear combination of elements in the rowspace is again in the rowspace, so in particular the linear combination given by $\mathbf{x}_\mathbf{r}^* - \hat{\mathbf{x}}_\mathbf{r}$ is in the rowspace. This contradicts the fact that this vector belongs to the nullspace. Therefore, there is no such $\mathbf{x}_\mathbf{r}^*$.

Problem 5: Expectation-Maximization. Let X be a random variable distributed according to $p_X(x)$ and Y be a random variable distributed according to $p_Y(y)$. Let $D_X = \{x_i\}_{i=1}^m$ be an i.i.d. sample from $p_X(x)$ and $D_Y = \{y_i\}_{i=1}^n$ be an i.i.d. sample from $p_Y(y)$. Let $D = D_X \cup D_Y$. Furthermore, define $p_X(x)$ and $p_Y(y)$ as follows:

$$p_X(x) = \alpha N(\mu_1, \sigma_1^2) + (1-\alpha)N(\mu_2, \sigma_2^2)$$

and

$$p_Y(y) = \beta N(\mu_1, \sigma_1^2) + (1-\beta)N(\mu_2, \sigma_2^2)$$

where $\alpha \in (0,1), \beta \in (0,1), \mu_1 \in \mathbb{R}, \mu_2 \in \mathbb{R}, \sigma_1 \in \mathbb{R}^+$ and $\sigma_2 \in \mathbb{R}^+$ are unknown parameters. $N(\mu, \sigma^2)$ is a univariate Gaussian distribution with mean μ and variance σ^2.

a) Derive update rules of an EM algorithm for estimating μ_1, μ_2, σ_1 and σ_2 based only on data set D_Y

b) Derive update rules of an EM algorithm for estimating $\alpha, \beta, \mu_1, \mu_2, \sigma_1$ and σ_2 based on data set D

Solution: In both cases the algorithm should follow the principle of maximizing the expected likelihood of complete data, i.e., if Z_i are hidden variables that indicate to which distribution observation i belongs, then we want to maximize

$$E_{\mathbf{Z}}[\log\, p(D, \mathbf{z}|\theta)|\theta^t]$$

by using the formula

$$\theta^{(t+1)} = \arg\max_\theta \{E_{\mathbf{Z}}[\log\, p(D, \mathbf{z}|\theta)|\theta^t]\}$$

The parameters θ will be the mean and variance of the distributions as well as the mixing coefficients.

a) In this case let us derive update rules based only on data set D_Y.

I will use slightly different notation for now. In what follows, $w_1 = \beta$ and $w_2 = (1-\beta)$. After deriving values for w_1 and w_2, I will convert back to β.
In this case $m = 2$ (two distributions), we get:

$$E_{\mathbf{Z}}[\log\, p(D, \mathbf{z}|\theta)|\theta^t] = \sum_{i=1}^n \Big[log(w_1 p(y_i|\theta_1))p_{Z_i}(1|y_i, \theta^{(t)}) + log(w_2 p(y_i|\theta_2))p_{Z_i}(2|y_i, \theta^{(t)})\Big] \qquad (*)$$

This is the equation we want to optimize, first with respect to w_1 and w_2, and then with respect to μ_1, μ_2, σ_1 and σ_2.

<u>For w_1 and w_2</u>: In this case we note that this is a constrain optimization since $w_1 + w_2 = 1$. So, by forming the Lagrangian with some constant c we get the following function, call it f:

$$f = \sum_{i=1}^n \Big[log(w_1 p(y_i|\theta_1))p_{Z_i}(1|y_i, \theta^{(t)}) + log(w_2 p(y_i|\theta_2))p_{Z_i}(2|y_i, \theta^{(t)})\Big] + c(w_1 + w_2) - 1$$

Now take partial derivatives and set to zero:

$$\frac{\partial f}{\partial w_1} = \sum_{i=1}^n \left[\frac{p_{Z_i}(1|y_i, \theta^{(t)})}{w_1}\right] + c = 0 \Longrightarrow w_1 = -\frac{\sum_{i=1}^n p_{Z_i}(1|y_i, \theta^{(t)})}{c}$$

$$\frac{\partial f}{\partial w_2} = \sum_{i=1}^n \left[\frac{p_{Z_i}(2|y_i, \theta^{(t)})}{w_2}\right] + c = 0 \Longrightarrow w_2 = -\frac{\sum_{i=1}^n p_{Z_i}(2|y_i, \theta^{(t)})}{c}$$

Also,

$$w_1 + w_2 = 1 \Longrightarrow -\frac{\sum_{i=1}^n p_{Z_i}(1|y_i, \theta^{(t)})}{c} - \frac{\sum_{i=1}^n p_{Z_i}(2|y_i, \theta^{(t)})}{c} = 1 \Longrightarrow$$

$$\frac{-\left[\sum_{i=1}^n p_{Z_i}(1|y_i, \theta^{(t)}) + p_{Z_i}(2|y_i, \theta^{(t)})\right]}{c} = 1 \Longrightarrow \frac{-\sum_{i=1}^n 1}{c} = 1 \Longrightarrow \frac{-n}{c} = 1 \Longrightarrow c = -n$$

So we can get rid c in w_1 and w_2 to get:

$$w_1 = \frac{\sum_{i=1}^{n} p_{Z_i}(1|y_i, \theta^{(t)})}{n} \qquad \text{and} \qquad w_2 = \frac{\sum_{i=1}^{n} p_{Z_i}(2|y_i, \theta^{(t)})}{n}$$

But then switch back to $\beta = w_1$ (and $1-\beta$ follows immediately) to get:

$$\beta = \frac{\sum_{i=1}^{n} p_{Z_i}(1|y_i, \theta^{(t)})}{n}$$

<u>For μ_1 and μ_2</u>: We will take the derivative of $(*)$ with respect to μ_k for $k = 1, 2$:

$$\begin{aligned}
\frac{\partial(*)}{\partial\mu_k} &= \frac{\partial}{\partial\mu_k}\left[\sum_{i=1}^{n} log(w_k p(y_i|\theta_k))p_{Z_i}(k|y_i, \theta^{(t)})\right] && \text{by definition of } (*) \\
&= \sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})\left[\frac{\partial}{\partial\mu_k} log(w_k p(y_i|\theta_k))\right] && \text{taking constants out} \\
&= \sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})\left[\frac{\partial}{\partial\mu_k}\{log(w_k) + log(p(y_i|\theta_k)))\}\right] && \text{properties of log} \\
&= \sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})\left[\frac{\partial}{\partial\mu_k} log(p(y_i|\theta_k))\right] \quad (**) && \text{since } log(w_k) \text{ is a constant w.r.t } \mu_k
\end{aligned}$$

where $p(y_i|\theta_k) = \frac{1}{\sigma_k\sqrt{2\pi}} e^{\left\{\frac{-(y_i-\mu_k)^2}{2\sigma_k^2}\right\}}$ and thus,

$$log(p(y_i|\theta_k)) = log\left[\frac{1}{\sigma_k\sqrt{2\pi}} e^{\left\{\frac{-(y_i-\mu_k)^2}{2\sigma_k^2}\right\}}\right] = log\left[\frac{1}{\sigma_k\sqrt{2\pi}}\right] - \frac{(y_i-\mu_k)^2}{2\sigma_k^2}$$

which means $\frac{\partial}{\partial\mu_k} log(p(y_i|\theta_k)) = \frac{\partial}{\partial\mu_k}\left\{log\left[\frac{1}{\sigma_k\sqrt{2\pi}}\right] - \frac{(y_i-\mu_k)^2}{2\sigma_k^2}\right\} = \frac{y_i-\mu_k}{\sigma_k^2}$

Replacing into $(**)$ and setting to zero:

$$\begin{aligned}
\frac{\partial(*)}{\partial\mu_k} &= \sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})\left[\frac{y_i-\mu_k}{\sigma_k^2}\right] && \text{replacing } \frac{\partial}{\partial\mu_k} log(p(y_i|\theta_k)) \text{ into } (**) \\
&= \frac{1}{\sigma_k^2}\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})(y_i-\mu_k) \\
&= \frac{1}{\sigma_k^2}\left[\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})y_i - \mu_k\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})\right] \\
&= 0 \qquad \Longrightarrow \text{(since } \sigma_k > 0) \\
\mu_k &= \frac{\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})y_i}{\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})}
\end{aligned}$$

Note that $\frac{\partial^2(**)}{\partial\mu_k^2} = \frac{1}{\sigma_k^2}\frac{\partial}{\partial\mu_k}\left[\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})y_i - \mu_k\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})\right] = -\frac{\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})}{\sigma_k^2} < 0.$

since the term in the numerator is the sum of probabilities and hence always positive and the term in the numerator is always positive since it is a square.

For σ_1 and σ_2: We will take the derivative of $(*)$ with respect to σ_k for $k = 1, 2$: Some of the work has already been done. Let us recap:

$$\frac{\partial(*)}{\partial\sigma_k} = \sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)}) \left[\frac{\partial}{\partial\sigma_k} log(p(y_i|\theta_k))\right] \quad \text{this is equation } (**) \text{, already computed}$$

Where, $\frac{\partial}{\partial\sigma_k} log(p(y_i|\theta_k)) = \frac{\partial}{\partial\sigma_k}\left\{log\left[\frac{1}{\sigma_k\sqrt{2\pi}}\right] - \frac{(y_i-\mu_k)^2}{2\sigma_k^2}\right\} = -\frac{1}{\sigma_k} + \frac{(y_i-\mu_k)^2}{\sigma_k^3}$

Replacing this into the equation above and setting to zero:

$$\begin{aligned}
\frac{\partial(*)}{\partial\sigma_k} &= \sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)}) \left[\frac{\partial}{\partial\sigma_k} log(p(y_i|\theta_k))\right] \\
&= \sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)}) \left[-\frac{1}{\sigma_k} + \frac{(y_i-\mu_k)^2}{\sigma_k^3}\right] \\
&= -\frac{1}{\sigma_k}\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)}) + \frac{1}{\sigma_k^3}\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})(y_i-\mu_k)^2 \\
&= 0 \qquad \Longrightarrow \text{multiplying by } \sigma_k
\end{aligned}$$

$$-\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)}) + \frac{1}{\sigma_k^2}\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})(y_i-\mu_k)^2 = 0 \Longrightarrow \sigma_k^2 = \frac{\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})(y_i-\mu_k)^2}{\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})}$$

An argument similar to the case for μ_k with the second derivative shows that this is a global maximum. (I will omit this argument here).

EM Algorithm: Following the class notes and the above derivation, the following is the EM Algorithm:

1. Initialize $\mu_k^{(0)}, \sigma_k^{(0)}$ for $k = 1, 2$ and $\beta^{(0)}$

2. Set $t = 0$.

3. Repeat until convergence

   (a) $p_{Z_i}(1|y_i, \theta^{(t)}) = \frac{\beta^{(t)}p(y_i|\mu_1, \sigma_1^2)}{\beta^{(t)}p(y_i|\mu_1, \sigma_1^2) + (1-\beta)^{(t)}p(y_i|\mu_2, \sigma_2^2)}$ and $p_{Z_i}(2|y_i, \theta^{(t)}) = 1 - p_{Z_i}(1|y_i, \theta^{(t)})$

   (b) $\beta^{(t+1)} = \frac{\sum_{i=1}^{n} p_{Z_i}(1|y_i, \theta^{(t)})}{n}$

   (c) $\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})y_i}{\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})}$

   (d) $(\sigma_k^2)^{(t+1)} = \frac{\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})(y_i-\mu_k)^2}{\sum_{i=1}^{n} p_{Z_i}(k|y_i, \theta^{(t)})}$

(e) $t = t + 1$

4. Report $\mu_k^{(t)}, (\sigma_k^2)^{(t)}$ and $\beta^{(t)}$ for $k = 1, 2$.

where $p(y_i|\mu_k, \sigma_k)$ is the pdf of a normal random variable with mean μ_k and variance σ_k^2

b) In this case let us derive update rules based on data set D. This data set contains data from both p_X and p_Y. I am going to use the notation:

$$w_1 = \frac{n}{n+m}\beta, \quad w_2 = \frac{n}{n+m}(1-\beta), \quad w_3 = \frac{m}{n+m}\alpha, \text{ and } w_4 = \frac{m}{n+m}(1-\alpha)$$

And hence:

$$w_1+w_2+w_3+w_4 = \frac{n}{n+m}\beta+\frac{n}{n+m}(1-\beta)+\frac{m}{n+m}\alpha+\frac{m}{n+m}(1-\alpha) = \frac{n\beta+n(1-\beta)+m\alpha+(1-m)\alpha}{n+m} = \frac{n+m}{n+m} = 1$$

So that we have a mixture of 4 distributions (which in reality reduces to only 2 distributions, but the calculations are the same) where each w is weighted by the number of points corresponding to the data set from which it came (either D_X with m points or D_Y with n points).

Now, the equation we want to optimize is:

$$\begin{aligned} E_{\mathbf{Z}}[\log p(D, \mathbf{z}|\theta)|\theta^t] &= \sum_{i=1}^{n+m} [log(w_1 p(y_i|\theta_1)) p_{Z_i}(1|y_i, \theta^{(t)}) + log(w_2 p(y_i|\theta_2)) p_{Z_i}(2|y_i, \theta^{(t)}) \\ &\quad + log(w_3 p(y_i|\theta_2)) p_{Z_i}(3|y_i, \theta^{(t)}) + log(w_4 p(y_i|\theta_2)) p_{Z_i}(4|y_i, \theta^{(t)})] \qquad (\circ) \end{aligned}$$

When we optimize equation ($\circ$) to find values for w_i, this essentially reduces to the computations done in part a), but we will have to account for w_3 and w_4. Form the Lagrangian, where c is a constant:

$$\begin{aligned} f &= \sum_{i=1}^{n+m} [log(w_1 p(y_i|\theta_1)) p_{Z_i}(1|y_i, \theta^{(t)}) + log(w_2 p(y_i|\theta_2)) p_{Z_i}(2|y_i, \theta^{(t)}) \\ &\quad + log(w_3 p(y_i|\theta_1)) p_{Z_i}(3|y_i, \theta^{(t)}) + log(w_4 p(y_i|\theta_2)) p_{Z_i}(4|y_i, \theta^{(t)})] + c(w_1+w_2+w_3+w_4) - 1 \end{aligned}$$

Now take partial derivatives and set to zero:

$$\frac{\partial f}{\partial w_1} = \sum_{i=1}^{n+m}\left[\frac{p_{Z_i}(1|y_i, \theta^{(t)})}{w_1}\right] + c = 0 \Longrightarrow w_1 = -\frac{\sum_{i=1}^{n+m} p_{Z_i}(1|y_i, \theta^{(t)})}{c}$$

$$\frac{\partial f}{\partial w_2} = \sum_{i=1}^{n+m}\left[\frac{p_{Z_i}(2|y_i, \theta^{(t)})}{w_2}\right] + c = 0 \Longrightarrow w_2 = -\frac{\sum_{i=1}^{n+m} p_{Z_i}(2|y_i, \theta^{(t)})}{c}$$

$$\frac{\partial f}{\partial w_3} = \sum_{i=1}^{n+m}\left[\frac{p_{Z_i}(3|y_i, \theta^{(t)})}{w_3}\right] + c = 0 \Longrightarrow w_3 = -\frac{\sum_{i=1}^{n+m} p_{Z_i}(3|y_i, \theta^{(t)})}{c}$$

$$\frac{\partial f}{\partial w_4} = \sum_{i=1}^{n+m}\left[\frac{p_{Z_i}(4|y_i, \theta^{(t)})}{w_4}\right] + c = 0 \Longrightarrow w_4 = -\frac{\sum_{i=1}^{n+m} p_{Z_i}(4|y_i, \theta^{(t)})}{c}$$

Also,

$$w_1+w_2+w_3+w_4 = 1 \Longrightarrow \frac{-\left[\sum_{i=1}^{n+m} p_{Z_i}(1|y_i, \theta^{(t)}) + p_{Z_i}(2|y_i, \theta^{(t)} + p_{Z_i}(3|y_i, \theta^{(t)} + p_{Z_i}(4|y_i, \theta^{(t)})\right]}{c} = 1 \Longrightarrow$$

$$-c = \left[\sum_{i=1}^{n+m} p_{Z_i}(1|y_i, \theta^{(t)}) + p_{Z_i}(2|y_i, \theta^{(t)}) + p_{Z_i}(3|y_i, \theta^{(t)}) + p_{Z_i}(4|y_i, \theta^{(t)})\right] = \sum_{j=1}^{4}\sum_{i=1}^{n+m} p_{Z_i}(j|y_i, \theta^{(t)})$$

So we can get rid c in w_1, w_2, w_3 and w_4 to get (for short, w_k for $k = 1, 2, 3, 4$):

$$w_k = \frac{\sum_{i=1}^{n+m} p_{Z_i}(k|y_i, \theta^{(t)})}{\sum_{j=1}^{4}\sum_{i=1}^{n+m} p_{Z_i}(j|y_i, \theta^{(t)})}$$

But then switch back to β from $w_1 = \frac{n}{n+m}\beta$, so $\beta = \frac{n+m}{n}w_1$ (and $1-\beta$ follows immediately) to get:

$$\beta = \left(\frac{n+m}{n}\right)\frac{\sum_{i=1}^{n+m} p_{Z_i}(1|y_i, \theta^{(t)})}{\sum_{j=1}^{4}\sum_{i=1}^{n+m} p_{Z_i}(j|y_i, \theta^{(t)})}$$

And switch back to α from $w_3 = \frac{m}{n+m}\alpha$ (and $1-\alpha$ follows immediately) to get:

$$\alpha = \left(\frac{n+m}{m}\right)\frac{\sum_{i=1}^{n+m} p_{Z_i}(3|y_i, \theta^{(t)})}{\sum_{j=1}^{4}\sum_{i=1}^{n+m} p_{Z_i}(j|y_i, \theta^{(t)})}$$

<u>For μ_1 and μ_2</u>: We will take the derivative of ($\circ$) with respect to μ_1 first (the other case is symmetrical):

$$\begin{aligned}
\frac{\partial(\circ)}{\partial\mu_1} &= \frac{\partial}{\partial\mu_1}\left[\sum_{i=1}^{n} log(w_1 p(y_i|\theta_1))p_{Z_i}(1|y_i,\theta^{(t)}) + log(w_3 p(y_i|\theta_1))p_{Z_i}(1|y_i,\theta^{(t)})\right] && \text{by definition of } (*)\\
&= \sum_{i=1}^{n+m} p_{Z_i}(1|y_i,\theta^{(t)})\left[\frac{\partial}{\partial\mu_1}log(w_1 p(y_i|\theta_1))\right] + \sum_{i=1}^{n+m} p_{Z_i}(3|y_i,\theta^{(t)})\left[\frac{\partial}{\partial\mu_1}log(w_3 p(y_i|\theta_1))\right] && \text{taking constants out}\\
&= \sum_{i=1}^{n+m} p_{Z_i}(1|y_i,\theta^{(t)})\left[\frac{\partial}{\partial\mu_1}log(p(y_i|\theta_1))\right] + \sum_{i=1}^{n+m} p_{Z_i}(3|y_i,\theta^{(t)})\left[\frac{\partial}{\partial\mu_1}log(p(y_i|\theta_1))\right] \quad (\circ\circ) && log(w_k) \text{ is a constan}
\end{aligned}$$

We already computed: $\frac{\partial}{\partial\mu_k}log(p(y_i|\theta_k)) = \frac{\partial}{\partial\mu_k}\left\{log\left[\frac{1}{\sigma_k\sqrt{2\pi}}\right] - \frac{(y_i-\mu_k)^2}{2\sigma_k^2}\right\} = \frac{y_i-\mu_k}{\sigma_k^2}$

Thus, replacing into ($\circ\circ$) and setting to zero:

$$\begin{aligned}
\frac{\partial(*)}{\partial\mu_k} &= \sum_{i=1}^{n+m} p_{Z_i}(1|y_i,\theta^{(t)})\left[\frac{y_i-\mu_1}{\sigma_1^2}\right] + \sum_{i=1}^{n+m} p_{Z_i}(3|y_i,\theta^{(t)})\left[\frac{y_i-\mu_1}{\sigma_1^2}\right] \quad \text{replacing } \frac{\partial}{\partial\mu_k}log(p(y_i|\theta_k)) \text{ into } (\circ\circ)\\
&= \frac{1}{\sigma_1^2}\left[\sum_{i=1}^{n+m} p_{Z_i}(1|y_i,\theta^{(t)})(y_i-\mu_1) + p_{Z_i}(3|y_i,\theta^{(t)})(y_i-\mu_1)\right]\\
&= \frac{1}{\sigma_1^2}\left[\sum_{i=1}^{n+m}(y_i-\mu_1)\left(p_{Z_i}(1|y_i,\theta^{(t)}) + p_{Z_i}(3|y_i,\theta^{(t)})\right)\right]\\
&= 0 \qquad \Longrightarrow \text{(since } \sigma_1 > 0)\\
\mu_1 &= \frac{\sum_{i=1}^{n+m}\left(p_{Z_i}(1|y_i,\theta^{(t)}) + p_{Z_i}(3|y_i,\theta^{(t)})\right)y_i}{\sum_{i=1}^{n+m} p_{Z_i}(1|y_i,\theta^{(t)}) + p_{Z_i}(3|y_i,\theta^{(t)})}
\end{aligned}$$

A very similar argument, which I will not write entirely to save space, shows that:

$$\mu_2 = \frac{\sum_{i=1}^{n+m}\left(p_{Z_i}(2|y_i,\theta^{(t)}) + p_{Z_i}(4|y_i,\theta^{(t)})\right)y_i}{\sum_{i=1}^{n+m} p_{Z_i}(2|y_i,\theta^{(t)}) + p_{Z_i}(4|y_i,\theta^{(t)})}$$

For σ_1 and σ_2: We will take the derivative of $(\circ)$ with respect to σ_1. Note that most of the work has already been done, so I am not going to write every detail here.

$$\frac{\partial(\circ)}{\partial\sigma_1} = \sum_{i=1}^{n+m} p_{Z_i}(1|y_i,\theta^{(t)})\left[\frac{\partial}{\partial\sigma_1}log(p(y_i|\theta_1))\right] + p_{Z_i}(3|y_i,\theta^{(t)})\left[\frac{\partial}{\partial\sigma_1}log(p(y_i|\theta_1))\right] \quad \text{this is equation } (\circ\circ).$$

Where, $\frac{\partial}{\partial\sigma_k}log(p(y_i|\theta_k)) = \frac{\partial}{\partial\sigma_k}\left\{log\left[\frac{1}{\sigma_k\sqrt{2\pi}}\right] - \frac{(y_i-\mu_k)^2}{2\sigma_k^2}\right\} = -\frac{1}{\sigma_k} + \frac{(y_i-\mu_k)^2}{\sigma_k^3}$

Replacing this into the equation above and setting to zero:

$$\begin{aligned}
\frac{\partial(*)}{\partial\sigma_1} &= \sum_{i=1}^{n+m} p_{Z_i}(1|y_i,\theta^{(t)})\left[\frac{\partial}{\partial\sigma_1}log(p(y_i|\theta_1))\right] + p_{Z_i}(3|y_i,\theta^{(t)})\left[\frac{\partial}{\partial\sigma_1}log(p(y_i|\theta_1))\right] \\
&= \sum_{i=1}^{n+m}\left(p_{Z_i}(1|y_i,\theta^{(t)}) + p_{Z_i}(3|y_i,\theta^{(t)})\right)\left[-\frac{1}{\sigma_1} + \frac{(y_i-\mu_1)^2}{\sigma_1^3}\right] \\
&= -\frac{1}{\sigma_1}\sum_{i=1}^{n+m}\left(p_{Z_i}(1|y_i,\theta^{(t)}) + p_{Z_i}(3|y_i,\theta^{(t)})\right)\left[1 - \frac{(y_i-\mu_1)^2}{\sigma_1^2}\right] \\
&= 0 \qquad \Longrightarrow \text{multiplying by } \sigma_1
\end{aligned}$$

$$\sigma_1^2 = \frac{\sum_{i=1}^{n+m}\left[p_{Z_i}(1|y_i,\theta^{(t)}) + p_{Z_i}(3|y_i,\theta^{(t)})\right](y_i-\mu_1)^2}{\sum_{i=1}^{n+m} p_{Z_i}(1|y_i,\theta^{(t)}) + p_{Z_i}(3|y_i,\theta^{(t)})}$$

Essentially the same argument shows that:

$$\sigma_2^2 = \frac{\sum_{i=1}^{n+m}\left[p_{Z_i}(2|y_i,\theta^{(t)}) + p_{Z_i}(4|y_i,\theta^{(t)})\right](y_i-\mu_2)^2}{\sum_{i=1}^{n+m} p_{Z_i}(2|y_i,\theta^{(t)}) + p_{Z_i}(4|y_i,\theta^{(t)})}$$

Finally, we have all the components we need for the EM algorithm:

EM Algorithm: Following the class notes and the above derivation, the following is the EM Algorithm:

1. Initialize $\mu_k^{(0)}, \sigma_k^{(0)}$ for $k = 1, 2$ and $\beta^{(0)}, \alpha^{(0)}$
2. Set $t = 0$.
3. Repeat until convergence
   (a) $p_{Z_i}(k|y_i,\theta^{(t)}) = \frac{w_k p(y_i|\theta_k)}{\sum_{i=1}^{4} w_i p(y_i|\theta_i)}$

   (b) $\beta^{(t+1)} = \left(\frac{n+m}{n}\right)\frac{\sum_{i=1}^{n+m} p_{Z_i}(1|y_i,\theta^{(t)})}{\sum_{j=1}^{4}\sum_{i=1}^{n+m} p_{Z_i}(j|y_i,\theta^{(t)})}$ and $\alpha^{(t+1)} = \left(\frac{n+m}{m}\right)\frac{\sum_{i=1}^{n+m} p_{Z_i}(3|y_i,\theta^{(t)})}{\sum_{j=1}^{4}\sum_{i=1}^{n+m} p_{Z_i}(j|y_i,\theta^{(t)})}$

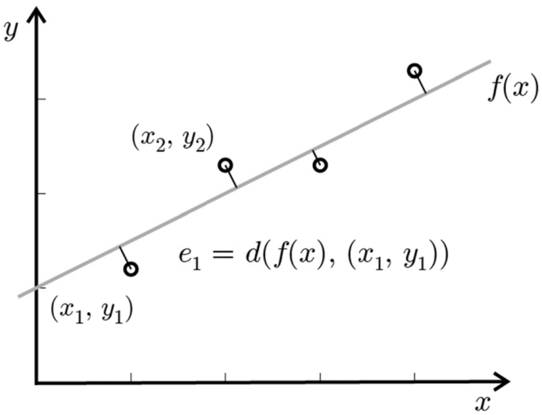
   (c) $\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n+m}\left(p_{Z_i}(k|y_i,\theta^{(t)}) + p_{Z_i}(k+2|y_i,\theta^{(t)})\right)y_i}{\sum_{i=1}^{n+m} p_{Z_i}(k|y_i,\theta^{(t)}) + p_{Z_i}(k+2|y_i,\theta^{(t)})}$

(d) $$(\sigma_k^2)^{(t+1)} = \frac{\sum_{i=1}^{n+m} \left[p_{Z_i}(k|y_i,\theta^{(t)}) + p_{Z_i}(k+2|y_i,\theta^{(t)})\right](y_i-\mu_1)^2}{\sum_{i=1}^{n+m} p_{Z_i}(k|y_i,\theta^{(t)}) + p_{Z_i}(k+2|y_i,\theta^{(t)})}$$

(e) $t = t+1$

4. Report $\mu_k^{(t)}, (\sigma_k^2)^{(t)}$ and $\beta^{(t)}, \alpha^{(t)}$ for $k = 1, 2$.

Problem 6: Consider the problem of linear regression in which the objective function is to minimize the sum of squared distances to the fitting line, as shown in the figure below. In the figure, $d(f(x),(x_0,y_0))$ represents the Euclidean distance from point (x_0,y_0) to the line $f(x)$. Formulate the optimization problem and solve it as far as you can make it. Assume you are given a data set $D = \{(x_i,y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}$, and $y_i \in \mathbb{R}$.



Solution: To solve this problem we first define the function $d(f(x),(x_0,y_0))$. As stated in [5.], given a point (x_0,y_0) and a line $ax+by+c=0$ with coefficients $a,b,c \in \mathbb{R}$, the perpendicular distance from the point to the line is given by:

$$d(f(x),(x_0,y_0)) = \sqrt{\frac{(ax_0+by_0+c)^2}{a^2+b^2}}$$

We can now define the problem of linear regression: suppose we are given data points $D = \{(x_i,y_i)\}_{i=1}^n$ where $x_i, y_i \in \mathbb{R}$. Let us hypothesize the fitting line to be $f(x_i) = w_0 + w_1 x_i$ or equivalently $w_1 x_i - f(x_i) + w_0 = 0$. The distance from the point (x_i,y_i) to this line is:

$$e_i = d(f(x_i),(x_i,y_i)) = \sqrt{\frac{(w_1x_i - y_i + w_0)^2}{w_1^2+(-1)^2}} \Longrightarrow e_i^2 = \frac{(w_1x_i - y_i + w_0)^2}{w_1^2+(-1)^2} = \frac{(w_1x_i - y_i + w_0)^2}{w_1^2+1}$$

Now we define the function $E(w_0,w_1)$ to be the sum of squared distances to the fitting line:

$$E(w_0,w_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n \frac{(w_1x_i - y_i + w_0)^2}{w_1^2+1} = \frac{1}{w_1^2+1}\sum_{i=1}^n (w_1x_i - y_i + w_0)^2$$

Minimization:

$$\frac{\partial E}{\partial w_0} = \frac{\partial}{\partial w_0}\left[\frac{1}{w_1^2+1}\sum_{i=1}^n (w_1x_i - y_i + w_0)^2\right] = \frac{2}{w_1^2+1}\sum_{i=1}^n (w_1x_i - y_i + w_0)$$

Setting $\frac{\partial E}{\partial w_0} = 0$ and noting that $w_1^2 + 1 > 0$, we get

$$\sum_{i=1}^n (w_1x_i - y_i + w_0) = 0 \Longrightarrow \sum_{i=1}^n (w_1x_i - y_i) + nw_0 = 0 \Longrightarrow \boxed{w_0 = \frac{\sum_{i=1}^n y_i - w_1x_i}{n}}$$

$$\frac{\partial E}{\partial w_1} = \frac{\partial}{\partial w_1}\left[\frac{1}{w_1^2+1}\sum_{i=1}^{n}(w_1x_i - y_i + w_0)^2\right] = \sum_{i=1}^{n}\frac{2(w_1x_i - y_i + w_0)(x_i + w_1(y_i - w_0))}{(w_1^2+1)^2}$$

Setting $\frac{\partial E}{\partial w_1} = 0$ and noting that $(w_1^2+1)^2 > 0$, we get

$$\boxed{\sum_{i=1}^{n}(w_1x_i - y_i + w_0)(x_i + w_1(y_i - w_0)) = 0}$$

Here we can replace w_0 from the first partial derivative into the above equation and solve for w_1. Next, replace the value found for w_1 into the equation for w_0 to obtain the optimal values for w_0 and w_1 in terms of the data only.

Even tough it is hard to get a close form solution for this problem, we could instead do an algorithm to approximate the solution. For example, we could do gradient descent: The first thing we need for this is the gradient ∇E, which we already computed:

$$\nabla E = \left(\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}\right) = \left(\frac{1}{w_1^2+1}\sum_{i=1}^{n}(w_1x_i - y_i + w_0)^2, \sum_{i=1}^{n}\frac{2(w_1x_i - y_i + w_0)(x_i + w_1(y_i - w_0))}{(w_1^2+1)^2}\right)$$

The algorithm would be:

<u>*Gradient Descent*</u>:

1. Initialize $w_0^{(0)}$ and $w_1^{(0)}$ (I suggest using the OLS solution)

2. Set $t = 0$.

3. Repeat until convergence

   (a) $w_0^{(t+1)} = w_0^{(t)} - \eta\frac{1}{(w_1^{(t)})^2+1}\sum_{i=1}^{n}(w_1^{(t)}x_i - y_i + w_0^{(t)})^2$

   (b) $w_1^{(t+1)} = w_1^{(t)} - \eta\sum_{i=1}^{n}\frac{2(w_1^{(t)}x_i - y_i + w_0^{(t)})(x_i + w_1^{(t)}(y_i - w_0^{(t)}))}{((w_1^{(t)})^2+1)^2}$

   (c) $t = t + 1$

4. Report $w_0^{(t)}, w_1^{(t)}$

where η is a parameter, usually a positive small number.

References

[1.] http://en.wikipedia.org/wiki/Gamma_distribution

[2.] http://en.wikipedia.org/wiki/Conjugate_prior

[3.] Essentials of Statistical Inference, G.A. Young and R. L. Smith

[4.] http://en.wikipedia.org/wiki/Beta_distribution

[5.] http://en.wikipedia.org/wiki/Distance_from_a_point_to_a_line