<u>H-W - 1</u> (Theoretical)

Problem d :

(i)
$$\int_0^{255} p(x)\, dx = \frac{\int_0^{255} f(x)\,dx}{\int_0^{255} f(x)\,dx} = 1$$

and
$$\int_0^{255} p(y)\, dy = \int_0^{125} \frac{1}{255}\, dy = 1$$

Hence,
$$p(x)\, dx = p(y)\,dy$$
$$= \frac{f(x)}{\int_0^{255} f(u)\,du}\, dx = \frac{1}{255}\, dy$$
$$\Rightarrow \frac{dy}{\cancel{dx}} = \frac{255 f(x)}{\int_0^{255} f(u)\,du}\, dx$$

Integrating both sides :-
$$y = \frac{1}{\int_0^{255} f(u)\,du} \int_0^{x} 255 f(x)\, dx = \frac{F(x-\mu) - F(0-\mu)}{F(255-\mu) - F(0-\mu)}$$

where F is the cumulative distribution function of a standard normal distribution, i.e. N(0 1)

$$\therefore \quad y = g(x) = \frac{F(x-127.5) - F(-127.5)}{F(127.5) - F(127.5)}$$

(ii)

$$P(x=x,\ y=y,\ z=z) = \begin{cases} 8xyz & \text{for } x,y,z \in [0,1] \\ 0 & \text{o/w.} \end{cases}$$

a) $P(x=x) = \int_z \int_y \underline{P(x=x,\ y=y,\ z=z)}\, dy\, dz$

$$= \int_{z=0}^{1} \int_{y=0}^{1} 8xyz\, dy\, dz = \int_z 8xz \cdot \frac{y^2}{2}\Big|_0^1 dz = 4x\frac{z^2}{2}\Big|_0^1$$

$$= \begin{cases} 2x & \text{for } x \in [0,1] \\ 0 & \text{o/w} \end{cases}$$

By Symmetric nature of probability function, we get

b) $P(y=y) = \begin{cases} 2y & \text{for } y \in [0,1] \\ 0 & \text{o/w.} \end{cases}$

c) $P(z=z) = \begin{cases} 2z & \text{for } z \in [0,1] \\ 0 & \text{o/w.} \end{cases}$

d) $E(xyz) = \int_z \int_y \int_x P(x,y,z) \cdot xyz\, dx\, dy\, dz$

$$= \int_{z=0}^{1} \int_{y=0}^{1} \int_{x=0}^{1} 8(xyz)^2\, dx\, dy\, dz = \frac{8}{27}$$

e) $P(x=x,\ y=y \mid z=z) = \dfrac{P(x=x,\ y=y,\ z=z)}{P(z=z)}$

$$= \begin{cases} \dfrac{8xyz}{2z} & \text{for } x,y \in [0,1] \\ 0 & \text{o/w.} \end{cases}$$

$$= \begin{cases} 4xy & \text{for } x, y \in [0,1] \\ 0 & \text{o/w} \end{cases}$$

Also, $P(x = x) \cdot P(y = y) = \begin{cases} 2x \cdot 2y & \text{for } x, y \in [0,1] \\ 0 & \text{o/w} \end{cases}$

$$= \begin{cases} 4xy & \text{for } x, y \in [0,1] \\ 0 & \text{o/w.} \end{cases}$$

Hence, x and y are conditionally independent given z.

Part e

(a) $MAP = \underset{\theta}{\text{argmax}} f(\theta|x)$ where $f(\theta|x)$ is the posterior distribution of θ.

$$f(\theta|x) = \frac{L(x|\theta)\ P(\theta)}{P(x)}$$

; where L is the likelihood function of x and $P(\theta)$ is the prior distribution of θ

$$\therefore \underset{\theta}{\text{argmax}}\ f(\theta|x) \equiv \max_{\theta} L(x|\theta) \cdot P(\theta).$$

$$= \underset{\theta}{\text{argmax}} \log\left(L(x|\theta) \cdot P(\theta)\right)$$

(since log is a 1-1 transformation & is increasing)

$\theta = (\mu, \Sigma)$.

Hence, $$\theta_{MAP} = \underset{\theta}{\text{argmax}} \log\left(\prod_{i=1}^{n} \left(\frac{1}{(2\pi)^{n/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)\right)\right)\right)$$

$$+ \log \frac{1}{(2\pi)^{n/2}} |\Sigma_0|^{-1/2} \exp\left(-\frac{1}{2}(\mu - \mu_0)^T \Sigma_0^{-1} (\mu - \mu_0)\right)$$

(i) $\therefore \mu_{MAP} = \max_{\mu} -\frac{1}{2}\sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) - \frac{1}{2}(\mu - \mu_0)^T \Sigma_0^{-1} (\mu - \mu_0)$

since Σ and Σ_0 are covariance matrices, they are positive semidefinite.

Hence the expression to be maximised ~~obtained~~ for μ_{MAP} is a convex function with a negative coefficient (A concave function). Hence differentiating w.r.t. μ will give us the Maximum value on equating with 0.

Differentiating the expression gives:-

$$0 = -\frac{1}{2}\left[\sum_{i=1}^{n}\left(2\mu^T\Sigma^{-1} - 2x_i^T\Sigma^{-1}\right) + 2\mu^T\Sigma_0^{-1} - 2\mu_0^T\Sigma_0^{-1}\right]$$

$$\Rightarrow \quad \mu^T(n\Sigma^{-1} + \Sigma_0^{-1}) = \mu_0^T\Sigma_0^{-1} + \sum_{1}^{n} x_i^T\Sigma^{-1}$$

Taking Transpose on both sides:-

$$\mu_{MAP} = (n\Sigma^{-1} + \Sigma_0^{-1})^{-1}\left(\Sigma_0^{-1}\mu_0 + \sum_{i=1}^{n}\Sigma^{-1}x_i\right)$$

(Σ & Σ_0 are Symmetric Matrices)

(ii) $$\Sigma_{MAP} = \underset{\Sigma}{\operatorname{argmax}} \sum_{i=1}^{n} -\frac{1}{2}\log(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

$$= \underset{\Sigma}{\operatorname{argmax}} -\frac{n}{2}\log(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

$$= \max_{\Sigma} -\frac{n}{2}\log(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{n}\text{Trace}\left(\Sigma^{-1}(x_i-\mu)(x_i-\mu)^T\right)$$

$$= \max_{\Sigma} -\frac{n}{2}\log(|\Sigma|) - \frac{1}{2}\text{Trace}\left(\Sigma^{-1}\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)\right)$$

The expression to Maximise is a concave function wrt. Σ^{-1}. Hence, differentiating wrt. Σ^{-1} and equating to 0 will give us the maximum value

(i) $\frac{d(|A|)}{dA} = A^{-T}$

(ii) $\frac{d\, Tr(AB)}{dA} = B^T$

$$\Rightarrow 0 = \frac{n}{2}(\Sigma^T) - \frac{1}{2}\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)^T$$

(These are symmetric)

Also, $$\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T = \sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T + n(\bar{x}-\mu)(\bar{x}-\mu)^T$$

where $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Hence, $$\Sigma_{MAP} = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)^T$$

b) (i) $$E(\mu_{MAP}) = E\,(n\Sigma^{-1}+\Sigma_0^{-1})^{-1}\left(\Sigma_0^{-1}\mu_0 + \sum_{i=1}^{n}\Sigma^{-1}x_i\right)$$

$$= (n\Sigma^{-1}+\Sigma_0^{-1})^{-1}\left(\Sigma_0^{-1}\mu_0 + \sum_{i=1}^{n}\Sigma^{-1}\mu\right)$$

$$= (n\Sigma^{-1}+\Sigma_0^{-1})^{-1}\left(\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\mu\right)$$

Hence, μ_{MAP} is a biased estimator unless

$\Sigma_0^{-1} = 0 \Rightarrow \Sigma_0$ has very large values.

(ii) $$E(\Sigma_{MAP}) = E\left(\frac{1}{N}\sum_{i=1}^{n}\left((x_i-\mu)(x_i-\mu)^T\right)\right)$$

$$= E\left(\frac{1}{N}\sum_{i=1}^{n}\left(x_i x_i^T - \mu x_i^T - x_i\mu^T + \mu\mu^T\right)\right)$$

$$= \frac{1}{N}\sum_{i=1}^{n}\left[E(x_i x_i^T) - E(\mu x_i^T) - E(x_i\mu^T) + E(\mu\mu^T)\right]$$

$$= \frac{1}{N}\sum_{i=1}^{n}\left[\Sigma - \mu\mu^T - \mu\mu^T + \mu\mu^T\right]$$

$$= \frac{1}{N}\sum_{i=1}^{n}(\Sigma - \mu\mu^T) = \Sigma - \mu\mu^T$$

Hence Σ_{MAP} is a~~n~~ ~~un~~biased estimator of Σ.

3. MLE of $\Sigma = \max_{\Sigma} f(x|\theta)$

$$= \max_{\Sigma} \log L(x|\theta)$$

$$= \max_{\Sigma} \sum_{i=1}^{n} \log \left(\frac{1}{(2\pi)^{n/2}} |\Sigma|^{-1/2} \exp\left((x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right) \right)$$

$$= \max_{\Sigma} -\frac{\cancel{n}}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^{n} \left((x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right)$$

This is the same formulation as Σ_{MAP}.

Hence, $\Sigma_{MAP} = \Sigma_{MLE} = \frac{1}{n} \sum_{i=1}^{n} \left((x_i - \mu)(x_i - \mu)^T \right)$

Also this implies that Σ_{MLE} is also a biased estimator of Σ

Part c :

Stochastic Gradient Descent Pseudo code :

for $i = 0, \ldots, N$ do :

$$A^i = X^T X - \sum_{j=0}^{i-1} \lambda^j d_j d_j^T$$

Initialise di randomly & let $t = 1$.

Choose k indexes randomly from length of di and store in a list smpl.

For (each k in smpl) do :

while ($t \leq T$ and stopping condition is not True) do:

Initialise a shared variable V with value same as that of dik.

$y_{new} = V - \eta \nabla_V (-V^2 A_i)$

update value of dik to y

$d_i = d_i / \| d_i \|$

$t = t + 1$

$\lambda^i = d_i^T X^T X d_i$