Sheet 08

EX1. $z \sim N(0, I)$ $\quad z \in \mathbb{R}^m$ $\quad m<d$ $\quad x \in \mathbb{R}^d$

$$X|z \sim N(Wz+\mu, \sigma^2 I) \qquad W(d\times m)$$

$$X = (x_n)_{n=1}^N \qquad Z = (z_n)_{n=1}^N$$

For iid data, the complete-data log likelihood:

$$\log p(X,Z|\theta) = \sum_{n=1}^N \{\log p(x_n|z_n;\theta) + \log p(z_n|\theta)\}$$

$\boxed{\text{here } \theta = (W, \mu, \sigma^2)}$

Gaussian density of $X \sim N(m, \Sigma)$:

$$p(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left[-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)\right]$$

if X is d-dim. then $\det(2\pi\Sigma) = (2\pi)^d \det(\Sigma)$

$$\log p(X,Z|\theta) = \sum_{n=1}^N \{\log p(x_n|z_n;\theta) + \log p(z_n|\theta)\} =$$

$$= \sum_{n=1}^N \{\log \mathcal{N}(Wz_n+\mu, \sigma^2 I_d) + \log \mathcal{N}(0, I_m)\} =$$

$$= \sum_{n=1}^N \Bigg\{ \log\left(\frac{1}{\sqrt{(2\pi)^d \underbrace{\det(\sigma^2 I)}_{=\sigma^{2d}}}} \exp\left[-\frac{1}{2}(x_n - Wz_n - \mu)^T [\sigma^2 I]^{-1}(x_n - Wz_n-\mu)\right]\right)$$

$$+ \log\left(\frac{1}{\sqrt{(2\pi)^m \underbrace{\det(I_m)}_{=1}}} \exp\left[-\frac{1}{2}(z_n - 0)^T[I]^{-1}(z_n - 0)\right]\right)\Bigg\} =$$

$$= \sum_{n=1}^N \Bigg\{ \log\left((2\pi)^{(m+d)\cdot(-\frac{1}{2})}\right) + \log\left[(\sigma^2)^{d\cdot(-\frac{1}{2})}\right] - \frac{1}{2}\cdot z_n^T z_n$$

$$- \frac{1}{2\cdot\sigma^2}(x_n - (Wz_n+\mu))^T(x_n - (Wz_n+\mu))\Bigg\} =$$

$$= -\frac{1}{2}(m+d)\cdot N \log(2\pi) - \frac{1}{2} d N \log(\sigma^2) + \sum_{i=1}^N \Bigg\{ -\frac{1}{2} z_n^T z_n - \frac{1}{2\sigma^2} z_n^T W^T W z_n$$

$$-\frac{1}{2\sigma^2}\left\{\left[(x_n-\mu)^T(x_n-\mu)\right] - (x_n-\mu)^T W z_n - z_n^T W^T (x_n - \mu)\right\}\Bigg\} =$$

$$= -\frac{1}{2}(m+d) N \log(2\pi) - d N \log(\sigma) + \sum_{i=1}^N \Bigg\{ -\frac{1}{2} z_n^T\left[I + \frac{1}{\sigma^2}W^T W\right] z_n$$

$$-\frac{1}{2}\left[(x_n-\mu)^T \frac{1}{\sigma^2} I (x_n-\mu)\right] + \frac{1}{2}\left[(x_n-\mu)^T \frac{1}{\sigma^2} W (z_n)\right] + \frac{1}{2}\left[(z_n)^T \frac{1}{\sigma^2} W^T (x_n - \mu)\right]\Bigg\} =$$

$$\cancel{\text{Rewrite: } [x_n-(Wz_n+\mu)]^T[x_n-(Wz_n+\mu)] =}$$

$$\cancel{= x_n^T x_n - (Wz_n+\mu)^T x_n - x_n^T(Wz_n+\mu) + (Wz_n+\mu)^T(Wz_n+\mu) =}$$

$$\cancel{= x_n^T x_n - (Wz_n)^T x_n - \mu^T x_n - x_n^T W z_n - x_n^T \mu + z_n^T W^T W z_n}$$

$$\cancel{+ z_n^T W^T \mu + \mu^T W z_n + \mu^T \mu}$$

Rewrite $[(x_n-\mu)-Wz_n]^T[(x_n-\mu)-Wz_n] =$

$$= (x_n-\mu)^T(x_n-\mu) - (x_n-\mu)^T W z_n - z_n^T W^T (x_n-\mu) + z_n^T W^T W z_n$$

$$= -\frac{1}{2}(m+d)N\log(2\pi) - dN\log(\sigma)$$

$$- \sum_{i=1}^{N} \frac{1}{2} \begin{bmatrix} z_n^T & (x_n-\mu)^T \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma^2}W^TW + I & -\frac{1}{\sigma^2}W^T \\ -\frac{1}{\sigma^2}W & \frac{1}{\sigma^2} I \end{bmatrix} \begin{bmatrix} z_n \\ x_n-\mu \end{bmatrix}$$

We can write the complete-data log likelihood function as an quadratic form. $z_n^T z_n$ is a scalar, so $z_n^T z_n = tr(z_n^T z_n) =$
$= tr(z_n z_n^T)$

$$= -\sum_{n=1}^{N} \left\{ \frac{m+d}{2}\log(2\pi) + \frac{d}{2}\log(\sigma^2) + \frac{1}{2} \cdot tr(z_n z_n^T) + \right.$$

$$\left. + \frac{1}{2\sigma^2}(x_n-\mu)^T(x_n-\mu) - \frac{1}{\sigma^2} z_n^T W^T (x_n-\mu) + \frac{1}{2\sigma^2} tr(W^T W z_n z_n^T) \right\}$$

$z_n^T W^T W z_n$ is a scalar, so $z_n^T W^T W z_n = tr(z_n^T W^T W z_n) =$
$= tr(W^T W z_n z_n^T)$

$z_n^T W^T (x_n-\mu)$ is a scalar, so $z_n^T W^T (x_n-\mu) = (x_n-\mu)^T W z_n$

In the E-step we take the expectation of $\log p(X, Z | \Theta)$ wrt. the distributions $p(z_n | x_n, \Theta)$.

(EX. 2a)

$$E_{z_n|x_n}\left[\log p(X, Z|\Theta)\right] = -\sum_{n=1}^{N} \left\{ \frac{m+d}{2}\log(2\pi) + \frac{d}{2}\log(\sigma^2) + \frac{1}{2} E_{z_n|x_n}\left[\mathrm{tr}(z_n z_n^T)\right] + \frac{1}{2\sigma^2}(x_n-\mu)^T(x_n-\mu) - \frac{1}{\sigma^2} E_{z_n|x_n}\left[z_n^T\right] W^T(x_n-\mu) + \frac{1}{2\sigma^2} E_{z_n|x_n}\left[\mathrm{tr}(W^T W z_n z_n^T)\right] \right\}$$

$$E_{z_n|x_n}\left[\mathrm{tr}(z_n z_n^T)\right] = \mathrm{tr}\left[E(z_n z_n^T | x_n)\right]$$

$$E_{z_n|x_n}\left[z_n^T\right] = E\left[z_n|x_n\right]^T$$

$$E_{z_n|x_n}\left[\mathrm{tr}(W^T W z_n z_n^T)\right] = \mathrm{tr}(W^T W E[z_n z_n^T | x_n])$$

therefore the $E_{z_n|x_n}[\log(p(X, Z|\Theta))]$ can be expressed using the variables $E[z_n|x_n]$ and $E[z_n z_n^T | x_n]$.

EX 2 a) Show that the E-step for the latent variable z_n can be expressed as follows:

(1.) $$E[z_n|x_n] = (\underset{m\times m}{W^T W} + \underset{m}{6^2 I})^{-1} \underset{m\times d}{W^T} (\underset{d}{x_n - \mu})$$

and (2.) $$E[z_n z_n^T|x_n] = 6^2 (W^T W + 6^2 I)^{-1} + E[z_n|x_n] \cdot E[z_n|x_n]^T$$

$z_n \sim N(0, I)$ $\quad z \in \mathbb{R}^m$ $\quad m < d$ $\quad , x \in \mathbb{R}^d$

$x_n|z_n \sim N(Wz + \mu, 6^2 I)$ $\quad W_{(d\times m)}$

Theorem: If $X_1 \sim N_r(\mu_1, \Sigma_{11})$ and $(X_2|X_1 = x_1) \sim N_{p-r}(Ax_1 + b, \Omega)$ where Ω does not depend on x_1, then

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_p(\mu, \Sigma)$$

where: $$\mu = \begin{pmatrix} \mu_1 \\ A\mu_1 + b \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{11} A^T \\ A\Sigma_{11} & \Omega + A\Sigma_{11} A^T \end{pmatrix}$$

We want to find the distribution of $\underline{x_n}$ $[p(x_n)]$.

Using the above theorem:

$X_1 := Z_n$ $\qquad X_2|X_1 := X_n|Z_n$

$\mu_1 = 0$ $\qquad A := W$

$\Sigma_{11} = I_m$ $\qquad b := \mu_m$

$\qquad \Omega := 6^2 I$

then $$\begin{pmatrix} Z_n \\ X_n \end{pmatrix} \sim N_{(d+m)} \left(\begin{bmatrix} 0_m \\ W \cdot 0_m + \mu_d \end{bmatrix}, \begin{bmatrix} I_m & I_m W^T \\ W I_d & 6^2 I_d + W I W^T \end{bmatrix} \right)$$

$$\begin{pmatrix} Z_n \\ X_n \end{pmatrix} \sim N_{(d+m)} \left(\begin{bmatrix} 0_m \\ \mu_d \end{bmatrix}, \begin{bmatrix} I_m & W^T \\ W & 6^2 I + WW^T \end{bmatrix} \right)$$

Therefore $X_n \sim N_d(\mu, 6^2 I + WW^T)$

Theorem : Assume $x \sim \mathcal{N}_x(\mu, \Sigma)$

$$x = \begin{bmatrix} x_a \\ x_b \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_a & \Sigma_c \\ \Sigma_c^T & \Sigma_b \end{bmatrix}$$

then

$$p(x_a | x_b) = \mathcal{N}_{x_a}(\hat{\mu}_a, \hat{\Sigma}_a)$$

$$\begin{cases} \hat{\mu}_a = \mu_a + \Sigma_c \Sigma_b^{-1} (x_b - \mu_b) \\ \hat{\Sigma}_a = \Sigma_a - \Sigma_c \Sigma_b^{-1} \Sigma_c^T \end{cases}$$

We want to find $E[z_n | x_n]$.

Let : $x_a := z_n$ $\qquad \begin{pmatrix} z_n \\ x_n \end{pmatrix} \sim \mathcal{N}_{d+m}\left(\begin{bmatrix} 0_m \\ \mu \end{bmatrix}, \begin{bmatrix} I_m & W^T \\ W & \sigma^2 I_d + WW^T \end{bmatrix} \right)$

$x_b := x_n$

then

$$\hat{\mu}_a = 0_m + \underset{m \times d}{W^T} \cdot [\underset{d \times d}{\sigma^2 I_d} + \underset{d \times d}{WW^T}]^{-1} (\underset{d}{x_n - \mu})$$

$$\hat{\Sigma}_a = I_m - W^T [\sigma^2 I_d + WW^T]^{-1} W$$

$$E[z_n | x_n] = W^T [\sigma^2 I + WW^T]^{-1} (x_n - \mu)$$

Using the identity (The Searle Set of Identities : Matrixcookbook p. 19)

$$A(I + BA)^{-1} = (I + AB)^{-1} A$$

$$E[z_n | x_n] = W^T [\sigma^2 I + WW^T]^{-1} (x_n - \mu) =$$

we can ignore σ^2 in the transformation, because it's just a scalar

$$= \frac{1}{\sigma^2} W^T [I + \frac{1}{\sigma^2} W W^T]^{-1} (x_n - \mu) =$$

$A := W^T \qquad B := \frac{1}{\sigma^2} W$

$$= \frac{1}{\sigma^2} [I + W^T \frac{1}{\sigma^2} W]^{-1} W^T (x_n - \mu) =$$

$$= [\sigma^2 I + W^T W]^{-1} W^T (x_n - \mu)$$

What was asked to be proved.

EX 2 a) 2.

From the previous theorem we have that

$$Z_n | X_n \sim N\left(\left[W^T [\sigma^2 I + WW^T]^{-1} (X_n - \mu) \right] ; \left[I_m - W^T [\sigma^2 I_d + WW^T]^{-1} W \right] \right)$$

Applying the Searle Identity

$$I - A(I + BA)^{-1} B = (I + AB)^{-1}$$

We get :

$$\mathrm{Var}(Z_n | X_n) = I_m - W^T [\sigma^2 I_d + WW^T]^{-1} W =$$

$$= I_m - \underbrace{\frac{1}{\sigma^2} W^T}_{=A} \Big[I_d + \underbrace{W}_{=B} \underbrace{\frac{1}{\sigma^2} W^T}_{A} \Big]^{-1} \underbrace{W}_{=B} =$$

$$= \left[I + \frac{1}{\sigma^2} W^T W \right]^{-1} = \sigma^2 (W^T W + \sigma^2 I)$$

~~From the law of total variance~~

$$\cancel{\mathrm{Var}(Z_n|X_n) = E[\mathrm{Var}(Z_n|X_n)]}$$

From the law of iterated expectations : $\mathrm{Var}(Z) = E(Z^2) - [E(X)]^2$

Then :

$$\mathrm{Var}(Z_n | X_n) = E[Z_n Z_n^T / X_n] - E[Z_n | X_n] \, E[Z_n | X_n]^T$$

So :

$$E[Z_n Z_n^T / X_n] = \mathrm{Var}[Z_n | X_n] + E[Z_n | X_n] \, E[Z_n | X_n]^T$$

$$E[Z_n Z_n^T / X_n] = \sigma^2 (W^T W + \sigma^2 I) + E[Z_n | X_n] \, E[Z_n | X_n]^T$$

what was to be shown.

EX. In the M-step, $E_{z_n|x_n}[\log p(X,Z|\theta)]$ is maximized wrt. W and σ^2, giving new parameter estimates.

$$\mathcal{L}_M = E_{z_n|x_n}[\log p(X|Z|\theta)] = -\sum_{n=1}^{N} \Big\{ \frac{m+d}{2}\log(2\pi) + \frac{d}{2}\log(\sigma^2) + \frac{1}{2}\,\mathrm{tr}\left(E[z_n z_n^T|x_n]\right) + \frac{1}{2\sigma^2}(x_n-\mu)^T(x_n-\mu)$$
$$- \frac{1}{\sigma^2} E[z_n|x_n]^T W^T (x_n-\mu) + \frac{1}{2\sigma^2}\mathrm{tr}\left(W^T W E[z_n z_n^T|x_n]\right)\Big\}$$

where
$$E[z_n|x_n] = (W^T W + \sigma^2 I)^{-1} W^T (x_n-\mu)$$
$$E[z_n z_n^T|x_n] = \sigma^2 (W^T W + \sigma^2 I)^{-1} + E[z_n|x_n]\cdot E[z_n|x_n]^T$$

If we consider the expectations computed in the E-step as a ground truth, we can focus on maximizing the observed-data log-likelihood, without substituting the expressions $E[z_n|x_n]$ and $E[z_n z_n^T|x_n]$ (treating them as given constants).

Then:
$$\frac{\partial \mathcal{L}_M}{\partial W} = \frac{\partial\left[\sum_{n=1}^{N} -\frac{1}{\sigma^2}E[z_n|x_n]^T W^T(x_n-\mu) + \frac{1}{2\sigma^2}\mathrm{tr}\left(W^T W E[z_n z_n^T|x_n]\right)\right]}{\partial W}$$

$$= -\sum_{n=1}^{N}\cdot\left(-\frac{1}{\sigma^2}\right)(x_n-\mu)\cdot E[z_n|x_n]^T - \sum_{n=1}^{N}\frac{1}{2\sigma^2}\left(W E[z_n z_n^T|x_n]^T + W E[z_n z_n^T|x_n]\right)$$

$$\boxed{\text{Since } \frac{\partial}{\partial X}\mathrm{tr}(X^T X B) = XB^T + XB \text{ (matrix Cookbook)}}$$

$$= \sum_{n=1}^{N}\frac{1}{\sigma^2}(x_n-\mu)E[z_n|x_n]^T - \sum_{n=1}^{N}\frac{1}{\cancel{2}\sigma^2}\left(\cancel{2}\cdot W E[z_n z_n^T|x_n]\right)$$

because $E[z_n z_n^T|x_n]$ is a symmetric matrix, so
$$E[z_n z_n^T|x_n] = E[z_n z_n^T|x_n]^T$$

setting the derivative $\frac{\partial \mathcal{L}_M}{\partial W}$ to zero, we get:
$$\cancel{\frac{1}{\sigma^2}}\sum_{n=1}^{N}(x_n-\mu)E[z_n|x_n]^T = \cancel{\frac{1}{\sigma^2}}\sum_{n=1}^{N} W E[z_n z_n^T|x_n])$$

then
$$\underline{\underline{W_{new}}} = \left[\sum_{n=1}^{N}(x_n-\mu)E[z_n|x_n]^T\right]\left[\sum_{n=1}^{N}E[z_n z_n^T|x_n]\right]^{-1}$$

Taking the derivative of $\mathcal{L}_M$ w.r.t. parameter σ^2:

$$\frac{\partial \mathcal{L}_M}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2}\Big[-\sum_{n=1}^{N}\Big\{\frac{d}{2}\log(\sigma^2) + \frac{1}{2\sigma^2}(x_n-\mu)^T(x_n-\mu) - \frac{1}{\sigma^2}E[z_n|x_n]^T W^T(x_n-\mu) + \frac{1}{2\sigma^2}tr(W^TWE[z_nz_n^T|x_n])\Big\}\Big]$$

$$= -\frac{Nd}{2}\cdot\frac{1}{\sigma^2} + (-1)\cdot\frac{1}{(\sigma^2)^2}\Big\{-\sum_{n=1}^{N}\Big\{\frac{1}{2}(x_n-\mu)^T(x_n-\mu) - E[z_n|x_n]^T W^T(x_n-\mu) + \frac{1}{2}tr(W^TWE[z_nz_n^T|x_n])\Big\}\Big\} =$$

$$= -\frac{Nd}{2}\cdot\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}\Big\{(x_n-\mu)^T(x_n-\mu) - 2E[z_n|x_n]^T W^T(x_n-\mu) + tr(W^TWE[z_nz_n^T|x_n])\Big\}$$

Setting the derivative $\frac{\partial \mathcal{L}_M}{\partial \sigma^2}$ to zero, we get:

$\frac{\partial \mathcal{L}_M}{\partial \sigma^2} = 0 \;/\cdot 2\sigma^2$ and rearranging the terms:

$$0 = -Nd + \frac{1}{\sigma^2}\sum_{n=1}^{N}\Big\{\|x_n-\mu\| - 2E[z_n|x_n]^T W^T(x_n-\mu) + tr\big(W^TWE[z_nz_n^T|x_n]\big)\Big\}$$

$$\sigma^2_{new} = \frac{1}{Nd}\sum_{n=1}^{N}\Big\{\|x_n-\mu\| - 2E[z_n|x_n]^T W^T(x_n-\mu) + tr\big(W^TWE[z_nz_n^T|x_n]\big)\Big\}$$

Therefore, to compute the updated value σ^2_{new} for the variance parameter, given the calculated values in the E-step:

$A_n := E[z_n|x_n]$ and $B_n := E[z_nz_n^T|x_n]$

we have to first calculate the new parameter W_{new} to subsequently calculate the new updated parameter σ^2_{new}, i.e.

$$W_{new} = \Big[\sum_{n=1}^{N}(x_n-\mu)A_n^T\Big]\Big[\sum_{n=1}^{N}B_n\Big]^{-1}$$

and

$$\sigma^2_{new} = \frac{1}{Nd}\sum_{n=1}^{N}\Big\{\|x_n-\mu\| - 2A_n^T W_{new}^T(x_n-\mu) + tr(W_{new}^T W_{new} B_n)\Big\}$$