



UNIVERSITY OF AMSTERDAM

University of Amsterdam

# Homework Assignment 3

Machine Learning I

2024

Pedro M. P. Curvo  
15713725

# Contents

1	Principal Component Analysis . . . . .	1
2	Probabilistic PCA - A general latent space distribution . . . . .	10
2.1	b) . . . . .	10
2.2	c) . . . . .	11
2.3	d) . . . . .	11
3	Mixtures of Experts . . . . .	12

# 1 Principal Component Analysis

a)

i)

Starting from the formalization of maximizing the scatter we have that:

$$\max \sum_{i=1}^n \|P\mathbf{x}_i - P\bar{\mathbf{x}}\|^2$$

Expanding the norm we have that:

$$\begin{aligned} \max \sum_{i=1}^n \|P\mathbf{x}_i - P\bar{\mathbf{x}}\|^2 &= \max \sum_{i=1}^n (P\mathbf{x}_i - P\bar{\mathbf{x}})^T (P\mathbf{x}_i - P\bar{\mathbf{x}}) \\ &= \max \sum_{i=1}^n (P(\mathbf{x}_i - \bar{\mathbf{x}}))^T P(\mathbf{x}_i - \bar{\mathbf{x}}) \\ &= \max \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^T P^T P (\mathbf{x}_i - \bar{\mathbf{x}}) \end{aligned}$$

Now, we need to consider two properties of the matrix  $P$ :

- $P$  is idempotent, i.e.  $P^n = P \quad \forall n \in \mathbb{N}$ , because projecting a vector twice is the same as projecting it once. If it is already in the subspace of the projection than another projection will not change it.
- $P$  is symmetric, i.e.  $P = P^T$

Now, we can rewrite the expression as:

$$\max \sum_{i=1}^n \|P\mathbf{x}_i - P\bar{\mathbf{x}}\|^2 = \max \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^T P (\mathbf{x}_i - \bar{\mathbf{x}})$$

Because,  $P^T P = PP = P$ . This is the same as the expression we wanted to show.

ii)

To prove that it is the same as  $\max \text{Tr}(S_1 P)$ , we first have to consider the Scatter matrix  $S_1$ :

$$S_1 = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T$$

This matrix is often used to estimate the covariance matrix and actually measures the scatter of the data by taking the outer product of the data points.

Now, we can rewrite the expression from the previous question as:

$$\begin{aligned}\max \sum_{i=1}^n \|P\mathbf{x}_i - P\bar{\mathbf{x}}\|^2 &= \max \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^T P (\mathbf{x}_i - \bar{\mathbf{x}}) \\ &= \max \sum_{i=1}^n \text{Tr} \left( (\mathbf{x}_i - \bar{\mathbf{x}})^T P (\mathbf{x}_i - \bar{\mathbf{x}}) \right) \quad \text{because the trace of a scalar is the scalar itself}\end{aligned}$$

Because the trace is invariant under cyclic permutations, i.e.  $\text{Tr}(ABC) = \text{Tr}(BCA) = \text{Tr}(CAB)$ , we can rewrite the expression as:

$$\begin{aligned}&= \max \sum_{i=1}^n \text{Tr} \left( P (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T \right) \\ &= \max \text{Tr} \left( P \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T \right) \\ &= \max \text{Tr} (PS_1) \quad \text{because } S_1 = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T\end{aligned}$$

This is the same as the expression we wanted to show.

**b)**

**i)**

Centering the data, which is done by subtracting the mean from each data point, is important for PCA because it removes the bias from the data. i.e., the bias introduced by the mean. This is important because the first principal component will always be the direction of the highest variance. If the data is not centered, the first principal component will be the direction that minimizes the distance to the data points, which is not what we want. This would also jeopardize the minimization of the projection error, because uncentered data would result in projections skewed towards the mean, leading to a higher projection error that does not reflect the actual variance of the data. In the plot above, we can see that the data is not centered, because the mean of the data is not at the origin, and we see two dashed lines that represent the first principal component, but passing through the origin. This shows that if the data is not centered the principal components will not align with the variance of the data, affected by the bias introduced by the mean and, therefore, affecting the projection error.

**ii)**

The reconstruction error is given, after centering, by:

$$\sum_{i=1}^n \|(\mathbf{x}_i - \bar{\mathbf{x}}) - P(\mathbf{x}_i - \bar{\mathbf{x}})\|^2$$

If we isolate  $(\mathbf{x}_i - \bar{\mathbf{x}})$  and expand the norm, we have:

$$\begin{aligned} \sum_{i=1}^n \|(\mathbf{x}_i - \bar{\mathbf{x}}) - P(\mathbf{x}_i - \bar{\mathbf{x}})\|^2 &= \sum_{i=1}^n \|(I - P)(\mathbf{x}_i - \bar{\mathbf{x}})\|^2 \\ &= \sum_{i=1}^n ((I - P)(\mathbf{x}_i - \bar{\mathbf{x}}))^T (I - P)(\mathbf{x}_i - \bar{\mathbf{x}}) \\ &= \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^T (I - P)^T (I - P)(\mathbf{x}_i - \bar{\mathbf{x}}) \end{aligned}$$

Now, since  $P$  is idempotent and symmetric, we have that  $(I - P)^T(I - P) = (I - P)(I - P) = I - P - P + P^2 = I - P$ .

Therefore, the expression becomes:

$$\min \sum_{i=1}^n \|(\mathbf{x}_i - \bar{\mathbf{x}}) - P(\mathbf{x}_i - \bar{\mathbf{x}})\|^2 = \min \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^T (I - P)(\mathbf{x}_i - \bar{\mathbf{x}})$$

As we wanted to show.

iii)

Similar to the previous question (a)ii),  $S_2$  is defined as:

$$S_2 = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T$$

And is the scatter matrix.

Following the same steps as before, we have that:

$$\begin{aligned} \min \sum_{i=1}^n \|(\mathbf{x}_i - \bar{\mathbf{x}}) - P(\mathbf{x}_i - \bar{\mathbf{x}})\|^2 &= \min \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^T (I - P) (\mathbf{x}_i - \bar{\mathbf{x}}) \\ &= \min \sum_{i=1}^n \text{Tr} \left( (\mathbf{x}_i - \bar{\mathbf{x}})^T (I - P) (\mathbf{x}_i - \bar{\mathbf{x}}) \right) \\ &= \min \text{Tr} \left( (I - P) \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T \right) \\ &= \min \text{Tr} (S_2(I - P)) \quad \text{because } S_2 = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T \end{aligned}$$

As we wanted to show.

c)

i)

Intuitively, a projection of a vector cannot have a higher length than the original vector, hence  $\|P\mathbf{x}_i\|^2 \leq \|\mathbf{x}_i\|^2$ .

First, let's expand the norms:

$$\begin{aligned}\|P\mathbf{y}\|^2 &= (P\mathbf{y})^T P\mathbf{y} \\ &= \mathbf{y}^T P^T P\mathbf{y} \\ &= \mathbf{y}^T P\mathbf{y}\end{aligned}$$

And:

$$\|\mathbf{y}\|^2 = \mathbf{y}^T \mathbf{y}$$

Now, since  $P$  is a projection matrix, it projects vectors onto a subspace, and this subspace cannot have a higher dimension than the original space. This also means it can only reduce or maintain the length of the vector. Formally the statement becomes:

$$P\mathbf{y} = \mathbf{y} \quad \forall \mathbf{y} \in \mathbb{R}^d \quad \text{if and only if } \mathbf{y} \in \text{Range}(P)$$

This gives that:

$$\begin{aligned}\|P\mathbf{y}\|^2 &= \mathbf{y}^T P\mathbf{y} \\ &= \mathbf{y}^T \mathbf{y} \quad \text{if } \mathbf{y} \in \text{Range}(P) \\ &\leq \mathbf{y}^T \mathbf{y} \\ &= \|\mathbf{y}\|^2\end{aligned}$$

As we wanted to show.

ii)

The preservation of the pair-wise distances as much as possible is given by:

$$\min \sum_{i=1}^n \sum_{j=1}^n \|\mathbf{x}_i - \mathbf{x}_j\|^2 - \|P\mathbf{x}_i - P\mathbf{x}_j\|^2$$

Expanding the norms, we have:

$$\begin{aligned}
\min \sum_{i=1}^n \sum_{j=1}^n \|\mathbf{x}_i - \mathbf{x}_j\|^2 - \|P\mathbf{x}_i - P\mathbf{x}_j\|^2 &= \min \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j) - (P\mathbf{x}_i - P\mathbf{x}_j)^T (P\mathbf{x}_i - P\mathbf{x}_j) \\
&= \min \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j) - (P(\mathbf{x}_i - \mathbf{x}_j))^T (P(\mathbf{x}_i - \mathbf{x}_j)) \\
&= \min \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j) - (\mathbf{x}_i - \mathbf{x}_j)^T P^T P (\mathbf{x}_i - \mathbf{x}_j) \\
&= \min \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j) - (\mathbf{x}_i - \mathbf{x}_j)^T P (\mathbf{x}_i - \mathbf{x}_j) \\
&= \min \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \mathbf{x}_j)^T ((\mathbf{x}_i - \mathbf{x}_j) - P(\mathbf{x}_i - \mathbf{x}_j)) \\
&= \min \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \mathbf{x}_j)^T (I - P) (\mathbf{x}_i - \mathbf{x}_j)
\end{aligned}$$

As we wanted to show.

iii)

First, let's consider the scatter matrix of pairwise differences  $S_3$ :

$$S_3 = \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \mathbf{x}_j) (\mathbf{x}_i - \mathbf{x}_j)^T$$

Contrary to the previous scatter matrices, this one considers the pairwise differences between all data points, and not the differences between the data points and the mean.

Now, using the cyclic and scalar properties of the trace, we have that:

$$\begin{aligned}
\min \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \mathbf{x}_j)^T (I - P) (\mathbf{x}_i - \mathbf{x}_j) &= \min \sum_{i=1}^n \sum_{j=1}^n \text{Tr} \left( (\mathbf{x}_i - \mathbf{x}_j)^T (I - P) (\mathbf{x}_i - \mathbf{x}_j) \right) \\
&= \min \text{Tr} \left( \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \mathbf{x}_j) (\mathbf{x}_i - \mathbf{x}_j)^T (I - P) \right) \\
&= \min \text{Tr} (S_3 (I - P)) \quad \text{because } S_3 = \sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \mathbf{x}_j) (\mathbf{x}_i - \mathbf{x}_j)^T
\end{aligned}$$

Now, we need to introduce a scaling factor due to the double sum in the scatter matrix  $S_3$ , because we are summing over all pairs  $(i, j)$ , and we are counting each pair twice, one for  $(i, j)$  and another for  $(j, i)$ . Also, we need to account for the scaling with the number of data points, since each individual datapoints is paired with all the others. Therefore, the scaling factor is given by  $2n$ , leading to:



$$\min \text{Tr} (S_3(I - P)) \rightarrow \min 2n \text{Tr} (S_3(I - P))$$

As we wanted to show.

**d)**

The three formulations presented in the previous questions can be reduced, as shown before, to:

$$\begin{aligned} \max \operatorname{Tr}(S_1 P), \\ \min \operatorname{Tr}(S_2(I - P)), \\ \min \operatorname{Tr}(S_3(I - P)), \end{aligned}$$

where  $P$  is the unknown projection matrix.

- The first formulation maximizes the scatter of the data projected onto the subspace. - The second formulation minimizes the reconstruction error of the data. - The third formulation minimizes the loss in pairwise distances of the data.

Since the trace operator returns a scalar and the identity matrix is symmetric, we can rewrite the second and third formulations as:

$$\begin{aligned} \max \operatorname{Tr}(S_2(P - I)), \\ \max \operatorname{Tr}(S_3(P - I)). \end{aligned}$$

Now, since we are maximizing the trace of a matrix, constant terms can be ignored. Specifically, in the second and third formulations,  $\operatorname{Tr}(-S_2 I)$  and  $\operatorname{Tr}(-S_3 I)$  are constant terms that do not affect the maximization with respect to  $P$ . Hence, we can equivalently rewrite them as:

$$\begin{aligned} \max \operatorname{Tr}(S_1 P), \\ \max \operatorname{Tr}(S_2 P), \\ \max \operatorname{Tr}(S_3 P). \end{aligned}$$

This shows that all three formulations reduce to maximizing the trace of a product of the projection matrix  $P$  and a scatter matrix  $S$  (which could be  $S_1$ ,  $S_2$ , or  $S_3$ ). The solution to all three formulations is therefore the same: the matrix  $P$  that maximizes the trace is the one that projects the data onto the subspace spanned by the top  $k$  eigenvectors of the covariance matrix (or scatter matrix).

Thus, the three formulations are equivalent, and they all lead to the same solution: the principal components found via PCA.

e)

The covariance matrix of the data is given by:

$$S = VDV^T$$

Where  $V$  is the matrix of eigenvectors and  $D$  is the diagonal matrix of eigenvalues. Those eigenvectors have the direction of the principal components, that is, the directions of the maximum variance of the data. Hence, the dotted lines in the plot represent the directions of the principal components and, thus, the directions of the eigenvectors of  $V$ , which are the columns of  $V$ .

Now, expressing the projection matrix as:

$$P = V_k V_k^T$$

Where  $V_k$  is the matrix of the first  $k$  eigenvectors of the covariance matrix.

Assuming  $k = 1$ , the projection matrix becomes:

$$P = V_1 V_1^T$$

Then, the projection of the data onto the first principal component by first removing the mean and adding the bias back, in order to reduce the error of the projection, is given by:

$$P\mathbf{x}_i = V_1 V_1^T (\mathbf{x}_i - \bar{\mathbf{x}}) + \bar{\mathbf{x}}$$

## 2 Probabilistic PCA - A general latent space distribution

a)

Since

$$p(x|z) = \mathcal{N}(Wz + \mu, \sigma^2 I)$$

Then we can say that:

$$\epsilon_x \sim \mathcal{N}(0, \sigma^2 I)$$

Now, since  $x = Wz + \mu + \epsilon_x$  and  $z = m + \epsilon_z$ , we can rewrite the expression as:

$$\begin{aligned} x &= W(m + \epsilon_z) + \mu + \epsilon_x \\ &= Wm + W\epsilon_z + \mu + \epsilon_x \end{aligned}$$

Now, since  $W\epsilon_z$  is a linear transformation of a Gaussian random variable, it is also Gaussian. Therefore,  $W\epsilon_z \sim \mathcal{N}(0, WW^T)$ .

With this, since the sum of two Gaussian random variables is also Gaussian, we have that:

$$x \sim \mathcal{N}(Wm + \mu, W\Sigma W^T + \sigma^2 I)$$

b)

To find the expectation of the variable  $x$  by taking into account the linearity of the expectation operator, we have that:

$$\begin{aligned} E[x] &= E[Wm + \mu + \epsilon_x + W\epsilon_z] \\ &= E[Wm] + E[\mu] + E[\epsilon_x] + E[W\epsilon_z] \\ &= Wm + \mu + E[\epsilon_x] + WE[\epsilon_z] \\ &= Wm + \mu \end{aligned}$$

This, because the expectancy of a constant is the constant itself in the case of  $Wm$  and  $\mu$ , and the expectancy of a Gaussian random variable is the mean of the Gaussian distribution, which is 0 in the case of  $\epsilon_x$ , since  $\epsilon_x \sim \mathcal{N}(0, \sigma^2 I)$  and in the case of  $W\epsilon_z$ , since  $\epsilon_z \sim \mathcal{N}(0, \Sigma)$ .

c)

To find the covariance of the variable  $x$ , we have that:

$$\begin{aligned} \text{Cov}[x] &= \text{Cov}[Wm + \mu + \epsilon_x + W\epsilon_z] \\ &= \text{Cov}[Wm] + \text{Cov}[\mu] + \text{Cov}[\epsilon_x] + \text{Cov}[W\epsilon_z] \\ &= W\text{Cov}[m]W^T + \text{Cov}[\mu] + \text{Cov}[\epsilon_x] + W\text{Cov}[\epsilon_z]W^T \\ &= \text{Cov}[\epsilon_x] + W\text{Cov}[\epsilon_z]W^T \quad \text{since } \text{Cov}[m] = \text{Cov}[\mu] = 0 \text{ because they are constants} \\ &= \sigma^2 I + W\Sigma W^T \quad \text{since } \text{Cov}[\epsilon_x] = \sigma^2 I \text{ and } \text{Cov}[\epsilon_z] = \Sigma \end{aligned}$$

This, because the covariance of a constant is 0, and the covariance of a Gaussian random variable is the covariance matrix of the Gaussian distribution,

d)

To match the previous expression:

$$x \sim \mathcal{N}(Wm + \mu, W\Sigma W^T + \sigma^2 I)$$

In the form:

$$x \sim \mathcal{N}(\tilde{\mu}, \tilde{W}\tilde{W}^T + \sigma^2 I)$$

Then, we have that:

$$\begin{aligned} \tilde{\mu} &= Wm + \mu \\ \tilde{W} &= W\Sigma^{1/2} \end{aligned}$$

The  $\tilde{\mu}$  is easy to check because it is the mean of the Gaussian distribution, and the  $\tilde{W}$  is the square root of the covariance matrix of the Gaussian distribution, since:

$$\begin{aligned} W\Sigma W^T + \sigma^2 I &= W\Sigma^{1/2}\Sigma^{1/2}W^T + \sigma^2 I \\ &= W\Sigma^{1/2}\Sigma^{1/2T}W^T + \sigma^2 I \quad \text{since } \Sigma \text{ is symmetric} \\ &= (W\Sigma^{1/2})(W\Sigma^{1/2})^T + \sigma^2 I \end{aligned}$$

### 3 Mixtures of Experts

a)

Considering that  $z_n$  is one-hot encoded, then we assign 1 to the index corresponding to the expert that is responsible for the data point  $x_n$ , i.e., the one with the highest probability, and 0 to the other indices. In resume, we need to find the index  $k$  that maximizes the probability  $p(z_n = k|x_n, \Phi) = \pi_{nk}$ .

This gives that:

$$z_n = \operatorname{argmax}_j \pi_{nj} = \begin{cases} 1, & \text{if } k = \operatorname{argmax}_j \pi_{nj} = \operatorname{argmax}_j \frac{\exp(\phi_j^T x_n)}{\sum_{l=1}^K \exp(\phi_l^T x_n)} \\ 0, & \text{otherwise} \end{cases}$$

b)

The likelihood of the data is given by:

$$\begin{aligned} p(y|X, \Theta, \Phi) &= \prod_{n=1}^N p(y_n|x_n, \Theta, \Phi) \quad \text{assuming i.i.d.} \\ &= \prod_{n=1}^N \sum_{k=1}^K p(y_n|z_n = k, x_n, \Theta) p(z_n = k|x_n, \Phi) \\ &= \prod_{n=1}^N \sum_{k=1}^K \pi_{nk} p(y_n|z_n = k, x_n, \Theta) \end{aligned}$$

Which, if we expand the terms, gives:

$$\begin{aligned} p(y|X, \Theta, \Phi) &= \prod_{n=1}^N \sum_{k=1}^K \frac{\exp(\phi_k^T x_n)}{\sum_{l=1}^K \exp(\phi_l^T x_n)} \exp(\theta_k^T x_n) \exp(-\exp(\theta_k^T x_n) y_n) \\ &= \prod_{n=1}^N \sum_{k=1}^K \frac{\exp(\phi_k^T x_n)}{\sum_{l=1}^K \exp(\phi_l^T x_n)} \exp(\theta_k^T x_n - \exp(\theta_k^T x_n) y_n) \end{aligned}$$

The log-likelihood, without the expanding terms, is given by:

$$\log p(y|X, \Theta, \Phi) = \sum_{n=1}^N \log \sum_{k=1}^K \pi_{nk} p(y_n|z_n = k, x_n, \Theta)$$

Expanding the terms, we have:

$$\log p(y|X, \Theta, \Phi) = \sum_{n=1}^N \log \sum_{k=1}^K \frac{\exp(\phi_k^T x_n)}{\sum_{l=1}^K \exp(\phi_l^T x_n)} \exp(\theta_k^T x_n - \exp(\theta_k^T x_n) y_n)$$

c)

The responsibility of the expert  $k$  for the data point  $x_n$  is given by:

$$\begin{aligned} r_{nk} &= p(z_n = k | x_n, \Theta, \Phi) \\ &= \frac{p(y_n | x_n, z_n = k, \Theta) p(z_n = k | x_n, \Phi)}{p(y_n | x_n, \Theta, \Phi)} \\ &= \frac{p(y_n | z_n = k, x_n, \Theta) p(z_n = k | x_n, \Phi)}{\sum_{j=1}^K p(y_n | z_n = j, x_n, \Theta) p(z_n = j | x_n, \Phi)} \\ &= \frac{\pi_{nk} p(y_n | z_n = k, x_n, \Theta)}{\sum_{j=1}^K \pi_{nj} p(y_n | z_n = j, x_n, \Theta)} \end{aligned}$$

By expanding the terms, we have:

$$r_{nk} = \frac{\frac{\exp(\phi_k^T x_n)}{\sum_{l=1}^K \exp(\phi_l^T x_n)} \exp(\theta_k^T x_n - \exp(\theta_k^T x_n) y_n)}{\sum_{j=1}^K \frac{\exp(\phi_j^T x_n)}{\sum_{l=1}^K \exp(\phi_l^T x_n)} \exp(\theta_j^T x_n - \exp(\theta_j^T x_n) y_n)}$$

Which can be simplified to:

$$r_{nk} = \frac{\exp(\phi_k^T x_n) \exp(\theta_k^T x_n - \exp(\theta_k^T x_n) y_n)}{\sum_{j=1}^K \exp(\phi_j^T x_n) \exp(\theta_j^T x_n - \exp(\theta_j^T x_n) y_n)}$$

d)

As shown before, the likelihood is given by:

$$p(y|X, \Theta, \Phi) = \prod_{n=1}^N \sum_{k=1}^K p(z_n = k | x_n, \Phi) p(y_n | z_n = k, x_n, \Theta)$$

By taking the hint into account to derivate the log-likelihood, we have that:

$$\begin{aligned}
\frac{\partial \log p(y|X, \Theta, \Phi)}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i} \sum_{n=1}^N \log \sum_{k=1}^K p(z_n = k|x_n, \Phi) p(y_n|z_n = k, x_n, \theta_k) \\
&= \sum_{n=1}^N \frac{\partial}{\partial \theta_i} \log \sum_{k=1}^K p(z_n = k|x_n, \Phi) p(y_n|z_n = k, x_n, \theta_k) \\
&= \sum_{n=1}^N \frac{1}{\sum_{k=1}^K p(z_n = k|x_n, \Phi) p(y_n|z_n = k, x_n, \theta_k)} \frac{\partial}{\partial \theta_i} \sum_{k=1}^K p(z_n = k|x_n, \Phi) p(y_n|z_n = k, x_n, \theta_k)
\end{aligned}$$

Now,  $p(z_n = k|x_n, \Phi)$  is given by:

$$p(z_n = k|x_n, \Phi) = \frac{\exp(\phi_k^T x_n)}{\sum_{l=1}^K \exp(\phi_l^T x_n)}$$

Therefore, it does not depend on  $\theta_i$ , giving that:

$$\begin{aligned}
\frac{\partial \log p(y|X, \Theta, \Phi)}{\partial \theta_i} &= \sum_{n=1}^N \frac{1}{\sum_{k=1}^K p(z_n = k|x_n, \Phi) p(y_n|z_n = k, x_n, \theta_k)} \frac{\partial}{\partial \theta_i} \sum_{k=1}^K p(z_n = k|x_n, \Phi) p(y_n|z_n = k, x_n, \theta_i) \\
&= \sum_{n=1}^N \frac{1}{\sum_{k=1}^K p(z_n = k|x_n, \Phi) p(y_n|z_n = k, x_n, \theta_k)} p(z_n = i|x_n, \Phi) \frac{\partial}{\partial \theta_i} p(y_n|z_n = i, x_n, \theta_i) \\
&= \sum_{n=1}^N r_{ni} \frac{\partial}{\partial \theta_i} p(y_n|z_n = i, x_n, \theta_i)
\end{aligned}$$

Now the derivative in respect to  $\phi_i$  is given by:

$$\begin{aligned}
\frac{\partial \log p(y|X, \Theta, \Phi)}{\partial \phi_i} &= \sum_{n=1}^N \frac{1}{\sum_{k=1}^K p(z_n = k|x_n, \Phi) p(y_n|z_n = k, x_n, \Theta)} \frac{\partial}{\partial \phi_i} \sum_{k=1}^K p(z_n = k|x_n, \Phi) p(y_n|z_n = k, x_n, \Theta) \\
&= \sum_{n=1}^N \frac{1}{\sum_{k=1}^K p(z_n = k|x_n, \Phi) p(y_n|z_n = k, x_n, \Theta)} \frac{\partial}{\partial \phi_i} \sum_{k=1}^K p(z_n = k|x_n, \Phi) p(y_n|z_n = k, x_n, \Theta) \\
&= \sum_{n=1}^N \frac{1}{\sum_{k=1}^K p(z_n = k|x_n, \Phi) p(y_n|z_n = k, x_n, \Theta)} p(y_n|z_n = i, x_n, \theta_i) \frac{\partial}{\partial \phi_i} p(z_n = i|x_n, \Phi)
\end{aligned}$$