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 11
3 Mixtures of Experts 13

1 Principal Component Analysis

a)

i)

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

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

Expanding the norm we have that:

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

Now, we need to consider two properties of the matrix $\boldsymbol{P}$:

- $\boldsymbol{P}$ is idempotent, i.e. $\boldsymbol{P}^n = \boldsymbol{P} \quad \forall n \in \mathbb{N}$, 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.
- $\boldsymbol{P}$ is symmetric, i.e. $\boldsymbol{P} = \boldsymbol{P}^T$

Now, we can rewrite the expression as:

$$\max \sum_{i=1}^{n} \|\boldsymbol{P}\boldsymbol{x_i} - \boldsymbol{P}\bar{\boldsymbol{x}}\|^2 = \max \sum_{i=1}^{n} \left(\boldsymbol{x_i} - \bar{\boldsymbol{x}}\right)^T \boldsymbol{P}\left(\boldsymbol{x_i} - \bar{\boldsymbol{x}}\right)$$

Because, $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$. This is the same as the expression we wanted to show.

ii)

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

$$\boldsymbol{S_1} = \sum_{i=1}^{n} \left(\boldsymbol{x_i} - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x_i} - \bar{\boldsymbol{x}}\right)^T$$

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:

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

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

$$= \max \sum_{i=1}^{n} Tr \left(\boldsymbol{P} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}) (\boldsymbol{x_i} - \bar{\boldsymbol{x}})^T \right)$$
$$= \max Tr \left(\boldsymbol{P} \sum_{i=1}^{n} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}) (\boldsymbol{x_i} - \bar{\boldsymbol{x}})^T \right)$$
$$= \max Tr (\boldsymbol{P}\boldsymbol{S_1}) \quad \text{because } \boldsymbol{S_1} = \sum_{i=1}^{n} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}) (\boldsymbol{x_i} - \bar{\boldsymbol{x}})^T$$

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. If the data is not centered, the first principal component may align more with the mean rather than with the covariance, leading to errors.

Besides that, when the data is not centered, the reconstruction error will be higher. This happens because the distance between the original data points and their projections onto the principal components will increase. The projection space is centered at the origin, while the original data may not be. This results in a constant bias in the distances, which cannot be corrected by the projection. Therefore, the reconstruction error needs to account for this bias, complicating the minimization process, specially because the bias is not constant in all directions, e.g, the bias in the direction of the first principal component might be twice as large as the bias in the direction of the second principal component. However, when the data is centered, the bias is removed and this bias distance is 0 in all directions, facilitating the minimization process and not introducing possible errors.

ii)

The reconstruction error is given, after centering, by:

$$\sum_{i=1}^{n} \left\| (\boldsymbol{x_i} - \bar{\boldsymbol{x}}) - \boldsymbol{P}\,(\boldsymbol{x_i} - \bar{\boldsymbol{x}}) \right\|^2$$

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

$$\begin{aligned}
\sum_{i=1}^{n} \left\| (\boldsymbol{x_i} - \bar{\boldsymbol{x}}) - \boldsymbol{P}\,(\boldsymbol{x_i} - \bar{\boldsymbol{x}}) \right\|^2 &= \sum_{i=1}^{n} \left\| (\boldsymbol{I} - \boldsymbol{P})\,(\boldsymbol{x_i} - \bar{\boldsymbol{x}}) \right\|^2 \\
&= \sum_{i=1}^{n} \left((\boldsymbol{I} - \boldsymbol{P})\,(\boldsymbol{x_i} - \bar{\boldsymbol{x}}) \right)^T (\boldsymbol{I} - \boldsymbol{P})\,(\boldsymbol{x_i} - \bar{\boldsymbol{x}}) \\
&= \sum_{i=1}^{n} (\boldsymbol{x_i} - \bar{\boldsymbol{x}})^T\,(\boldsymbol{I} - \boldsymbol{P})^T (\boldsymbol{I} - \boldsymbol{P})\,(\boldsymbol{x_i} - \bar{\boldsymbol{x}})
\end{aligned}$$

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

Therefore, the expression becomes:

$$\min \sum_{i=1}^{n} \left\| (\boldsymbol{x_i} - \bar{\boldsymbol{x}}) - \boldsymbol{P}\,(\boldsymbol{x_i} - \bar{\boldsymbol{x}}) \right\|^2 = \min \sum_{i=1}^{n} (\boldsymbol{x_i} - \bar{\boldsymbol{x}})^T\,(\boldsymbol{I} - \boldsymbol{P})\,(\boldsymbol{x_i} - \bar{\boldsymbol{x}})$$

As we wanted to show.

iii)

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

$$\boldsymbol{S_2} = \sum_{i=1}^{n} (\boldsymbol{x_i} - \boldsymbol{x_i}) (\boldsymbol{x_i} - \boldsymbol{x_i})^T$$

And is the scatter matrix.
Following the same steps as before, we have that:

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

As we wanted to show.

c)

i)

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

We can prove it by decomposing the vector $\boldsymbol{y}$ into two components, one that is in the subspace of the projection and another that is orthogonal to it.

$$\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y} + (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$$

Meaning that:

$$\begin{aligned}
\|\boldsymbol{y}\|^2 &= \|\boldsymbol{P}\boldsymbol{y} + (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}\|^2 \\
&= \|\boldsymbol{P}\boldsymbol{y}\|^2 + \|(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}\|^2 + 2\boldsymbol{y}^T\boldsymbol{P}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} \\
&= \|\boldsymbol{P}\boldsymbol{y}\|^2 + \|(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}\|^2 + 2\boldsymbol{y}^T\boldsymbol{P}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} \\
&= \|\boldsymbol{P}\boldsymbol{y}\|^2 + \|(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}\|^2 + 2\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{P}^2\boldsymbol{y} \\
&= \|\boldsymbol{P}\boldsymbol{y}\|^2 + \|(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}\|^2 + 2\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} \\
&= \|\boldsymbol{P}\boldsymbol{y}\|^2 + \|(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}\|^2
\end{aligned}$$

The second step could also be done by saying that $\boldsymbol{P} \perp (\boldsymbol{I} - \boldsymbol{P}) \rightarrow \boldsymbol{P}^T(\boldsymbol{I} - \boldsymbol{P}) = 0$.

Now, a norm is always positive, hence:

$$\begin{aligned}
\|\boldsymbol{P}\boldsymbol{y}\|^2 &= \|\boldsymbol{y}\|^2 - \|(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}\|^2 \\
&\leq \|\boldsymbol{y}\|^2 \quad \text{as we wanted to show}
\end{aligned}$$

Hence, the contraction property is proved.

ii)

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

$$\min \sum_{i=1}^{n}\sum_{j=1}^{n} \|\boldsymbol{x_i} - \boldsymbol{x_j}\|^2 - \|\boldsymbol{P}\boldsymbol{x_i} - \boldsymbol{P}\boldsymbol{x_j}\|^2$$

Expanding the norms, we have:

$$\begin{aligned}
\min \sum_{i=1}^{n}\sum_{j=1}^{n} \|\boldsymbol{x_i}-\boldsymbol{x_j}\|^2 - \|\boldsymbol{P}\boldsymbol{x_i}-\boldsymbol{P}\boldsymbol{x_j}\|^2 &= \min \sum_{i=1}^{n}\sum_{j=1}^{n} (\boldsymbol{x_i}-\boldsymbol{x_j})^T(\boldsymbol{x_i}-\boldsymbol{x_j}) - (\boldsymbol{P}\boldsymbol{x_i}-\boldsymbol{P}\boldsymbol{x_j})^T(\boldsymbol{P}\boldsymbol{x_i}-\boldsymbol{P}\boldsymbol{x_j}) \\
&= \min \sum_{i=1}^{n}\sum_{j=1}^{n} (\boldsymbol{x_i}-\boldsymbol{x_j})^T(\boldsymbol{x_i}-\boldsymbol{x_j}) - (\boldsymbol{P}(\boldsymbol{x_i}-\boldsymbol{x_j}))^T(\boldsymbol{P}(\boldsymbol{x_i}-\boldsymbol{x_j})) \\
&= \min \sum_{i=1}^{n}\sum_{j=1}^{n} (\boldsymbol{x_i}-\boldsymbol{x_j})^T(\boldsymbol{x_i}-\boldsymbol{x_j}) - (\boldsymbol{x_i}-\boldsymbol{x_j})^T\boldsymbol{P}^T\boldsymbol{P}(\boldsymbol{x_i}-\boldsymbol{x_j}) \\
&= \min \sum_{i=1}^{n}\sum_{j=1}^{n} (\boldsymbol{x_i}-\boldsymbol{x_j})^T(\boldsymbol{x_i}-\boldsymbol{x_j}) - (\boldsymbol{x_i}-\boldsymbol{x_j})^T\boldsymbol{P}(\boldsymbol{x_i}-\boldsymbol{x_j}) \\
&= \min \sum_{i=1}^{n}\sum_{j=1}^{n} (\boldsymbol{x_i}-\boldsymbol{x_j})^T((\boldsymbol{x_i}-\boldsymbol{x_j}) - \boldsymbol{P}(\boldsymbol{x_i}-\boldsymbol{x_j})) \\
&= \min \sum_{i=1}^{n}\sum_{j=1}^{n} (\boldsymbol{x_i}-\boldsymbol{x_j})^T(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{x_i}-\boldsymbol{x_j}) \\
&= \min \sum_{i,j=1}^{n} (\boldsymbol{x_i}-\boldsymbol{x_j})^T(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{x_i}-\boldsymbol{x_j})
\end{aligned}$$

As we wanted to show.

Note: Again, we used the idempotent and symmetric properties of the projection matrix $\boldsymbol{P}$.

iii)

First, let's consider the scatter matrix of pairwise differences $\boldsymbol{S_3}$:

$$\boldsymbol{S_3} = \sum_{i=1}^{n}\sum_{j=1}^{n} (\boldsymbol{x_i}-\boldsymbol{x_j})(\boldsymbol{x_i}-\boldsymbol{x_j})^T$$

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:

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

Now, we need to introduce a scaling factor due to the double sum in the scatter matrix $\boldsymbol{S_3}$, 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 scalling 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:

$$\min Tr\left(\boldsymbol{S_3}(\boldsymbol{I}-\boldsymbol{P})\right) \rightarrow \min 2nTr\left(\boldsymbol{S_3}(\boldsymbol{I}-\boldsymbol{P})\right)$$

As we wanted to show.

d)

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

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

where $\boldsymbol{P}$ 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:

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

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

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

This shows that all three formulations reduce to maximizing the trace of a product of the projection matrix $\boldsymbol{P}$ and a scatter matrix $\boldsymbol{S}$ (which could be $\boldsymbol{S_1}$, $\boldsymbol{S_2}$, or $\boldsymbol{S_3}$). The solution to all three formulations is therefore the same: the matrix $\boldsymbol{P}$ 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:

$$\boldsymbol{S} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T$$

Where $\boldsymbol{V}$ is the matrix of eigenvectors and $\boldsymbol{D}$ 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. In the plot, the yellow line corresponds to the direction of V_1 and the brown line corresponds to the direction of V_2.

Now, for any given space that is spanned by the column space of A, the projection matrix P is given by:

$$\boldsymbol{P} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$$

Since, in the case of PCA, the matrix A is the matrix of eigenvectors, the projection matrix P is given by:

$$\boldsymbol{P} = \boldsymbol{V}_k\boldsymbol{V}_k^T$$

Where $\boldsymbol{V}_k$ is the matrix of the first k eigenvectors of the covariance matrix. This is because, the eigenvectors are orthogonal to each other, hence:

$$\boldsymbol{P} = \boldsymbol{V}_k(\boldsymbol{V}_k^T\boldsymbol{V}_k)^{-1}\boldsymbol{V}_k^T = \boldsymbol{V}_k(\boldsymbol{I})^{-1}\boldsymbol{V}_k^T = \boldsymbol{V}_k\boldsymbol{V}_k^T$$

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

$$\boldsymbol{P} = \boldsymbol{V}_1\boldsymbol{V}_1^T$$

Then, the projection of the data onto the first principal component by adding the bias back is given by:

$$\boldsymbol{P}\boldsymbol{x_i} = \boldsymbol{V}_1\boldsymbol{V}_1^T\boldsymbol{x_i} + \bar{\boldsymbol{x}}$$

Not considering the bias, the projection of the data onto the first principal component is given by:

$$\boldsymbol{P}\boldsymbol{x_i} = \boldsymbol{V}_1\boldsymbol{V}_1^T\boldsymbol{x_i}$$

If we consider the correction of the bias, the projection will be a parallel line to the yellow dotted line, in which the bias is the distance between the origin and the mean of the data. If we do not consider the

correction of the bias, the points will be projected onto the yellow dotted line.

Plotting the projection of the data onto the first principal component, we have:

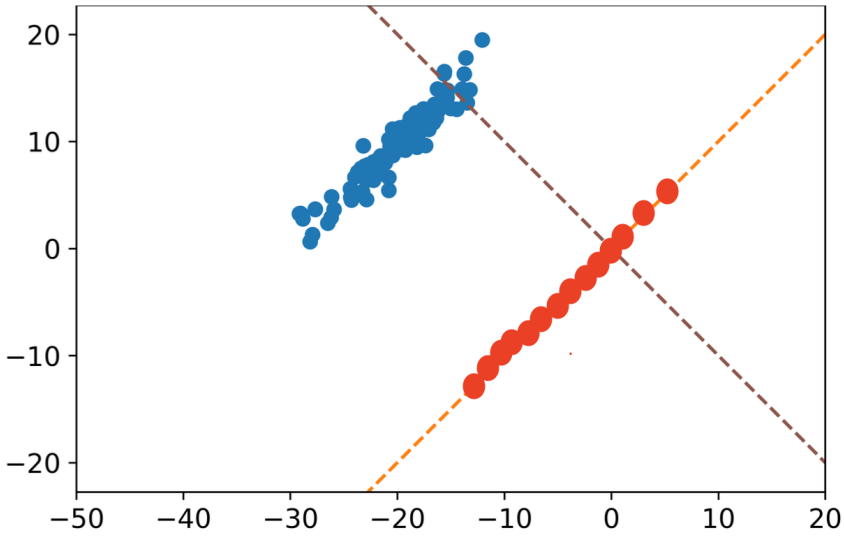

Fig. 1: Projection of the data onto the first principal component

2 Probabilistic PCA - A general latent space distribution

a)

Since

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

Then we can say that:

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

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

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

Now, since $\boldsymbol{W}\epsilon_z$ is a linear transformation of a Gaussian random variable, it is also Gaussian. Therefore, $\boldsymbol{W}\epsilon_z \sim \mathcal{N}(0, \boldsymbol{W}\Sigma\boldsymbol{W}^T)$ (this will be shown in the next question).

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

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

Note: We can ignore the $\boldsymbol{W}(\boldsymbol{m}+\boldsymbol{\mu}$ because they are constants and do not affect the type of distribution of $\boldsymbol{x}$.

b)

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

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

This, because the expectancy of a constant is the constant itself in the case of $\boldsymbol{W}\boldsymbol{m}$ and $\boldsymbol{\mu}$, and the

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

c)

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

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

This, because the covariance of a Gaussian random variable is the covariance matrix of the Gaussian distribution.

d)

To match the previous expression for the General Gaussian prior:

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

In the form:

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

We have that:

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

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

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

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 $\boldsymbol{x_n}$, 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 $p(z_n = k|\boldsymbol{x_n}, \Phi) = \pi_{nk}$.

This gives that:

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

b)

The likelihood of the data is given by:

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

Which, if we expand the terms, gives:

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

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

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

Expanding the terms, we have:

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

c)

The responsability of the expert i for the data point $\boldsymbol{x_n}$ is given by:

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

By expanding the terms, we have:

$$r_{ni} = \frac{\frac{\exp(\boldsymbol{\phi_i^T x_n})}{\sum_{l=1}^{K} \exp(\boldsymbol{\phi_l^T x_n})} \exp(\boldsymbol{\theta_i^T x_n} - \exp(\boldsymbol{\theta_i^T x_n}) \boldsymbol{y_n})}{\sum_{j=1}^{K} \frac{\exp(\phi_j^T \boldsymbol{x_n})}{\sum_{l=1}^{K} \exp(\boldsymbol{\phi_l^T x_n})} \exp(\theta_j^T \boldsymbol{x_n} - \exp(\theta_j^T \boldsymbol{x_n}) \boldsymbol{y_n})}$$

Which can be simplified to:

$$r_{ni} = \frac{\exp(\boldsymbol{\phi_i^T x_n}) \exp(\boldsymbol{\theta_i^T x_n} - \exp(\boldsymbol{\theta_i^T x_n}) \boldsymbol{y_n})}{\sum_{j=1}^{K} \exp(\phi_j^T \boldsymbol{x_n}) \exp(\theta_j^T \boldsymbol{x_n} - \exp(\theta_j^T \boldsymbol{x_n}) \boldsymbol{y_n})}$$

d)

As shown before, the likelihood is given by:

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

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

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

Since, $p(\boldsymbol{z_n}=k|\boldsymbol{x_n},\Phi)$ does not depend on $\boldsymbol{\theta_i}$, we have that:

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

Now, taking the derivative in respect of $\boldsymbol{\phi_i}$:

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

Now, since $p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\boldsymbol{\theta_k})$ does not depend on $\boldsymbol{\phi_i}$, we have that:

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

Note: The responsabilities only appear when taking the explicitly derivation of the probabilities, since it will depende on the values of the probabilities. This is be shown in the next question.

e)

Now, inserting the explicit expressions for the probabilities, we have that:

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

Since π_{nk} does not depend on $\boldsymbol{\theta_i}$, we have that:

$$\begin{aligned}\frac{\partial \log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\Phi})}{\partial \boldsymbol{\theta_i}} &= \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_{nk} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\boldsymbol{\Theta})} \sum_{k=1}^{K} \pi_{nk} \frac{\partial}{\partial \boldsymbol{\theta_i}} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\Theta) \\ &= \sum_{n=1}^{N} \frac{\pi_{nk}}{\sum_{k=1}^{K} \pi_{nk} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\boldsymbol{\Theta})} \sum_{k=1}^{K} \frac{\partial}{\partial \boldsymbol{\theta_i}} \left(\lambda \exp(-\lambda \boldsymbol{y_n})\right) \\ &= \sum_{n=1}^{N} \frac{\pi_{nk}}{\sum_{k=1}^{K} \pi_{nk} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\boldsymbol{\Theta})} \sum_{k=1}^{K} \frac{\partial}{\partial \lambda} \left(\lambda \exp(-\lambda \boldsymbol{y_n})\right) \frac{\partial \lambda}{\partial \boldsymbol{\theta_i}} \quad \text{by the chain rule} \\ &= \sum_{n=1}^{N} \frac{\pi_{nk}}{\sum_{k=1}^{K} \pi_{nk} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\boldsymbol{\Theta})} \sum_{k=1}^{K} (\exp(-\lambda \boldsymbol{y_n}) - \lambda \boldsymbol{y_n} \exp(-\lambda \boldsymbol{y_n})) \frac{\partial \lambda}{\partial \boldsymbol{\theta_i}} \\ &= \sum_{n=1}^{N} \frac{\pi_{nk}}{\sum_{k=1}^{K} \pi_{nk} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\boldsymbol{\Theta})} \sum_{k=1}^{K} (\exp(-\lambda \boldsymbol{y_n}) - \lambda \boldsymbol{y_n} \exp(-\lambda \boldsymbol{y_n})) \frac{\partial}{\partial \boldsymbol{\theta_i}} \exp(\boldsymbol{\theta_k}^T \boldsymbol{x_n}) \\ &= \sum_{n=1}^{N} \frac{\pi_{nk}}{\sum_{k=1}^{K} \pi_{nk} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\boldsymbol{\Theta})} \sum_{k=1}^{K} (\exp(-\lambda \boldsymbol{y_n}) - \lambda \boldsymbol{y_n} \exp(-\lambda \boldsymbol{y_n})) \exp(\boldsymbol{\theta_k}^T \boldsymbol{x_n}) \boldsymbol{x_n}^T \delta_{ik} \\ &= \sum_{n=1}^{N} \frac{\pi_{ni}}{\sum_{k=1}^{K} \pi_{nk} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\boldsymbol{\Theta})} (\exp(-\lambda \boldsymbol{y_n}) - \lambda \boldsymbol{y_n} \exp(-\lambda \boldsymbol{y_n})) \exp(\boldsymbol{\theta_i}^T \boldsymbol{x_n}) \boldsymbol{x_n}^T \\ &= \sum_{n=1}^{N} \frac{\pi_{ni}}{\sum_{k=1}^{K} \pi_{nk} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\boldsymbol{\Theta})} \lambda \exp(-\lambda \boldsymbol{y_n}) (\mathbb{1} - \lambda \boldsymbol{y_n}) \boldsymbol{x_n}^T \\ &= \sum_{n=1}^{N} \frac{\pi_{ni} p(\boldsymbol{y_n}|\boldsymbol{z_n}=i,\boldsymbol{x_n},\boldsymbol{\theta_i})}{\sum_{k=1}^{K} \pi_{nk} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\boldsymbol{\Theta})} (\mathbb{1} - \lambda \boldsymbol{y_n}) \boldsymbol{x_n}^T \quad \text{since } p(\boldsymbol{y_n}|\boldsymbol{z_n}=i,\boldsymbol{x_n},\boldsymbol{\theta_i}) = \lambda \exp(-\lambda \boldsymbol{y_n}) \\ &= \sum_{n=1}^{N} r_{ni} \boldsymbol{x_n}^T (\mathbb{1} - \lambda \boldsymbol{y_n})\end{aligned}$$

Now, for $\boldsymbol{\phi_i}$:

$$
\begin{aligned}
\frac{\partial \log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\Phi})}{\partial \boldsymbol{\phi_i}} &= \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_{nk} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\phi)} \frac{\partial}{\partial \boldsymbol{\phi_i}} \sum_{k=1}^{K} p(\boldsymbol{z_n}=k|\boldsymbol{x_n},\Phi) p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\boldsymbol{\phi_k}) \\
&= \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_{nk} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\Theta)} \frac{\partial}{\partial \boldsymbol{\phi_i}} \sum_{k=1}^{K} \pi_{nk} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\Theta) \\
&= \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_{nk} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\Theta)} \sum_{k=1}^{K} \frac{\partial}{\partial \boldsymbol{\phi_i}} \pi_{nk} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\Theta) \\
&= \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_{nk} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\Theta)} \sum_{k=1}^{K} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\Theta) \frac{\partial}{\partial \boldsymbol{\phi_i}} \pi_{nk} \\
&= \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_{nk} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\Theta)} \sum_{k=1}^{K} p(\boldsymbol{y_n}|\boldsymbol{z_n}=k,\boldsymbol{x_n},\Theta) \frac{\partial}{\partial \boldsymbol{\phi_i}} \frac{\exp(\boldsymbol{\phi_k^T x_n})}{\sum_{l=1}^{K} \exp(\boldsymbol{\phi_l^T x_n})} \\
&= \sum_{n=1}^{N} \frac{p(\boldsymbol{y_n}|z_n=k,\boldsymbol{x_n},\Theta)\pi_{ni}(1-\pi_{ni})\boldsymbol{x_n^T} - \sum_{k\neq i}^{K} p(\boldsymbol{y_n}|z_n=k,\boldsymbol{x_n},\Theta)\pi_{nk}\pi_{ni}\boldsymbol{x_n^T}}{\sum_{k=1}^{K} \pi_{nk} p(\boldsymbol{y_n}|z_n=k,\boldsymbol{x_n},\Theta)} \\
&= \sum_{n=1}^{N} \left[r_{ni}(1-\pi_{ni})\boldsymbol{x_n^T} - \sum_{k\neq i}^{K} r_{nk}\pi_{ni}\boldsymbol{x_n^T} \right] \\
&= \sum_{n=1}^{N} \left[r_{ni}(1-\pi_{ni})\boldsymbol{x_n^T} - \pi_{ni}\boldsymbol{x_n^T} \sum_{k\neq i}^{K} r_{nk} \right] \\
&= \sum_{n=1}^{N} \left[r_{ni}(1-\pi_{ni})\boldsymbol{x_n^T} - \pi_{ni}\boldsymbol{x_n^T}(1-r_{ni}) \right] \quad \text{since} \sum_{k=1}^{K} r_{nk} = 1 \rightarrow \sum_{k\neq i}^{K} r_{nk} = 1 - r_{ni} \\
&= \sum_{n=1}^{N} \left[r_{ni} - \pi_{ni} \right] \boldsymbol{x_n^T}
\end{aligned}
$$

f)

Writing down an iterative algorithm that maximizes the log-probability by jointly optimizing the parameters Θ and Φ:

Algorithm 1 Joint optimization of Θ and Φ in Mixtures of Experts

Input: Data X, labels Y, parameters Θ, Φ, learning rate α, tolerance ϵ, maximum iterations
Output: Updated parameters Θ, Φ
Initialize parameters Θ and Φ
while not converged **do**
E-step: Compute responsibilities r_{nk}:

$$r_{nk} = \frac{p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_k})\, \pi_{nk}}{\sum_{j=1}^{K} p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta}_j)\, \pi_{nj}}$$

where $\pi_{nk} = p(\boldsymbol{z_n} = k|\boldsymbol{x_n}, \Phi) = \frac{\exp(\boldsymbol{\phi_k^T x_n})}{\sum_{j=1}^{K} \exp(\phi_j^T \boldsymbol{x_n})}$
M-step: Update Θ and Φ using the gradients:

$$\theta_i^T \leftarrow \theta_i^T + \alpha \sum_{n=1}^{N} r_{ni} \boldsymbol{x_n^T} (\mathbb{1} - \lambda \boldsymbol{y_n})$$

$$\phi_i^T \leftarrow \phi_i^T + \alpha \sum_{n=1}^{N} [r_{ni} - \pi_{ni}]\, \boldsymbol{x_n^T}$$

Check for convergence:
Compute the change in log-likelihood:

$$\Delta\mathcal{L} = \left|\mathcal{L}^{(t+1)} - \mathcal{L}^{(t)}\right|$$

if $\Delta\mathcal{L} < \epsilon$ **or** maximum iterations reached **then**
Terminate: The algorithm has converged when the log-likelihood improvement is smaller than a predefined threshold ϵ, indicating that further updates provide negligible improvement.
Return optimized parameters Θ and Φ.
else
Continue to the next iteration by updating $t \leftarrow t + 1$.
end if
end while

g)

If instead of having a single expert, we have multiple experts, then the final prediction $\hat{\boldsymbol{y}}$ is given by the weighted sum of the predictions of each expert, where the weights are given by the responsibilities r_{nk}:

$$\hat{\boldsymbol{y}} = \sum_{k=1}^{K} r_{nk} \hat{\boldsymbol{y}}_{nk}$$

This $\hat{\boldsymbol{y}}_{nk}$ is the prediction of the expert k for the data point $\boldsymbol{x_n}$, which can be sampled from the distribution of expert k.