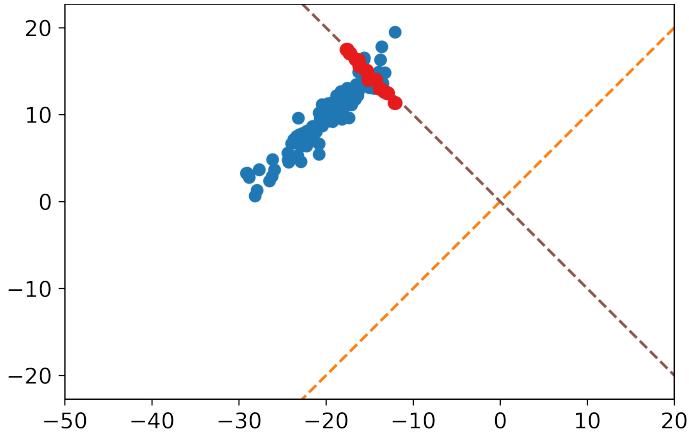


## 1 Principal Component Analysis (9 points)



There are many ways of defining PCA. Here we will focus on three of them. Remember that we are given  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$  and we want to find an orthogonal projection matrix  $P \in \mathbb{R}^{d \times d}$  of rank  $k$ . The goal of PCA is to get projections ( $P\mathbf{x}_1, \dots, P\mathbf{x}_n \in \mathbb{R}^d$ ) that “retain most of the original information”. This is of course very hand wavy, so let us formalize it in three different ways.

(A) Maximize the scatter.

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

(B) After centering, minimize the reconstruction error.

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

(C) Preserve the pair-wise distances as much as possible. A projection matrix is a contraction, so  $\|P\mathbf{y}\|^2 \leq \|\mathbf{y}\|^2$ . Thus, all pair-wise distances  $\|\mathbf{x}_i - \mathbf{x}_j\|^2$  will become smaller after the projection, and we want to make this gap as small as possible.

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

You will prove that the three formulations are equivalent

(a) For formulation (A)...

(i) Prove that it is equivalent to

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

[1 point]

(ii) Prove that it is equivalent to

$$\max \text{Tr}(S_1 P).$$

What is  $S_1$  in this case? [1 point]

(b) For formulation (B)...

(i) Why do we want to center the data before minimizing the reconstruction error? Use the scatter plot above to make your point. [1 point]

(ii) Prove that it is equivalent to

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

[1 points]

(iii) Prove that it is equivalent to

$$\min \text{Tr}(S_2(I - P)).$$

What is  $S_2$  in this case? [0.5 points]

(c) For formalization (C)...

(i) (Bonus): prove the contraction property of projection matrices. [1 point]

(ii) Prove that it is equivalent to

$$\min \sum_{i,j=1}^n (\mathbf{x}_i - \mathbf{x}_j)^T (I - P) (\mathbf{x}_i - \mathbf{x}_j)$$

[0.5 points]

(iii) Prove that it is equivalent to

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

What is  $S_3$  in this case? [1.5 points]

(d) Finally, show that the three formulations are equivalent. [1 point]

(e) The solution of this problem relies on the eigendecomposition of the covariance matrix  $S$ :

$$S = VDV'$$

- What is the relationship between  $V$  and the dotted lines on the scatter plot above? [0.5 points]
- Can you express  $P$  in terms of  $V$ ? (you can just state the final solution, no need to prove anything here) [0.5 points]
- Plot the projected points on the scatter plot above, assuming  $k = 1$ . [0.5 points]

## 2 Probabilistic PCA - A general latent space distribution (5 points)

Principal Component Analysis (PCA) is an often used technique for dimensionality reduction. In this approach, we linearly project data onto the subspace of lower dimensionality. However, this approach can be extended to be more general by formulating a latent variable model named probabilistic PCA. In this approach, the PCA can be expressed as the maximum likelihood solution probabilistic PCA. There are multiple advantages of the probabilistic PCA over the conventional PCA. To name a few:

- We can associate a likelihood function to the probabilistic PCA which allows a direct comparison with other probabilistic density models.
- Probabilistic PCA can be used to model class-conditional densities and can thus be used in classification problems,
- We can run the model generatively to provide samples from the modeled distribution.

Probabilistic PCA is an example of the linear-Gaussian framework, where both marginal and conditional distributions are Gaussian. We first define a latent variable  $\mathbf{z}$ , corresponding to the principal-component subspace. We can then define a prior distribution  $p(\mathbf{z})$  over the latent variable  $\mathbf{z}$ , and also the conditional distribution  $p(\mathbf{x}|\mathbf{z})$ , which is given by

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

The prior distribution over  $\mathbf{z}$  is usually given by a zero-mean unit-covariance Gaussian

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}).$$

In this case, the marginal distribution  $p(\mathbf{x})$  is also a Gaussian, and is given by

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{WW}^T + \sigma^2 \mathbf{I}).$$

However, suppose we replace the zero-mean and the unit-covariance latent space distribution by a general Gaussian distribution of the form

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{m}, \boldsymbol{\Sigma}).$$

In this problem, we wish to show that by redefining parameters of the model, this assumption on the prior leads to an identical model for the marginal distribution  $p(\mathbf{x})$  over the observed variables for any valid choice of  $\mathbf{m}$  and  $\boldsymbol{\Sigma}$ . We will derive this result in multiple steps.

- (a) We can express the random variable  $\mathbf{z}$  as  $\mathbf{z} = \mathbf{m} + \boldsymbol{\epsilon}_z$ , with noise  $\boldsymbol{\epsilon}_z \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ . Write out a similar expression for the variable  $\mathbf{x} = \mathbf{Wz} + \boldsymbol{\mu} + \boldsymbol{\epsilon}_x$ , and explain

what distribution variable  $\mathbf{x}$  follows. From which distribution is  $\epsilon_x$  sampled from? [1 point]

(b) Find the expectation value of the variable  $\mathbf{x}$  using the linearity property of the expected value. [1 point]

(c) Find the covariance of the variable  $\mathbf{x}$  using the definition of the covariance  $\text{cov}[\mathbf{x}, \mathbf{x}]$ . [2 point]

*Hint:* The following identity might be helpful:  $\text{Var}[\mathbf{AY}] = \mathbf{A}\text{Var}[\mathbf{Y}]\mathbf{A}^T$ , where  $\mathbf{A}$  is a matrix, and  $\mathbf{Y}$  is a random variable.

(d) To show that using the general Gaussian prior still leads to an identical model  $p(\mathbf{x})$ , we have to be able to write the distribution  $p(\mathbf{x})$  in the form  $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\tilde{\boldsymbol{\mu}}, \tilde{\mathbf{W}}\tilde{\mathbf{W}}^T + \sigma^2\mathbf{I})$ . Find the appropriate expressions for  $\tilde{\boldsymbol{\mu}}$  and  $\tilde{\mathbf{W}}$ . [1 point]

## Third assignment in Machine learning 1 – 2025 – Paper 1

### 3 Mixture of experts (8 points)

In class, you discussed and were introduced to mixture models as a way to perform unsupervised learning tasks, *e.g.* clustering. Mixture models can be similarly used for supervised learning tasks. In this question, we will discuss and explore the Mixtures of Experts (MoEs), a model that softly partitions the input space and learns a supervised model for each area.

Consider that you have  $K$  expert models available in order to model a specific dataset of  $N$  data points  $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$ , where  $\mathbf{x}_n \in \mathbb{R}^D$  and  $y_n$  corresponds to the ground truth label for  $\mathbf{x}_n$ . Let  $z_n$  denote a categorical random variable for the data point  $\mathbf{x}_n$  that denotes which of the  $K$  expert models is 'active'. Thus, if  $z_n = k$ , then expert model  $k$  is active for datapoint  $\mathbf{x}_n$ , meaning that model  $k$  provides the prediction for datapoint  $\mathbf{x}_n$ . Furthermore, let  $\mathbf{z}_n$  correspond to a one-hot encoding of  $z_n$ . The strength of an MoE approach is that it trains expert models that each specialise in a specific portion of the data.

Then, let  $\boldsymbol{\Theta}$  be a matrix in  $\mathbb{R}^{D \times K}$  that contains the  $D$ -dimensional column vector of parameters for each expert. We will assume that each  $y_i$  is a continuous random variable at the  $[0, \infty)$  interval and is distributed according to an exponential distribution with a rate  $\lambda > 0$ . Given the aforementioned assumptions, each expert  $k \in K$  has the following linear predictive model:

$$p(y_n | \mathbf{x}_n, \mathbf{z}_n, \boldsymbol{\Theta}) = p(y_n | \mathbf{x}_n, \boldsymbol{\theta}_k = \boldsymbol{\Theta}\mathbf{z}_n) = \text{Exponential}(y_n | \lambda = \exp(\boldsymbol{\theta}_k^T \mathbf{x}_n))$$

where  $\mathbf{z}_n$  is the one-hot-encoded vector representation of the categorical variable  $z_n$  and

$$\text{Exponential}(y | \lambda) = \lambda \exp(-\lambda y) \text{ for } y \geq 0.$$

The flexibility of MoEs stems from the fact that there is a "routing" mechanism which determines how relevant each of the  $K$  experts is for a specific datapoint  $\mathbf{x}_n$ . As in this case we have a discrete set of  $K$  experts, a simple linear routing mechanism is the following:

$$p(z_n = k | \mathbf{x}_n, \boldsymbol{\Phi}) = \pi_{nk} = \frac{\exp(\boldsymbol{\phi}_k^T \mathbf{x}_n)}{\sum_j \exp(\boldsymbol{\phi}_j^T \mathbf{x}_n)}$$

where  $\boldsymbol{\Phi}$  is a matrix in  $\mathbb{R}^{D \times K}$  that contains all of the parameters of the routing function, i.e.  $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_K]$ .

- (a) Note that the output of the routing function falls between 0 and 1. Thus, we need to construct our vector  $\mathbf{z}_n$  based on these values.

Write down the formula to decide each element  $z_{nk}$  of  $\mathbf{z}_n$ , assuming you want the most relevant expert to be active for datapoint  $\mathbf{x}_n$ . [0.5 points]

*Hint: remember that  $\mathbf{z}_n$  is one-hot encoded.*

As a-priori we have no information about which of the experts is responsible for generating a particular prediction, we have to marginalize over all possible experts in order to compute the likelihood of an observed point. With this information answer the following questions:

- (b) Write down the likelihood  $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\Theta}, \boldsymbol{\Phi})$  and the log-likelihood of the entire dataset. [1.5 points]
- (c) Write down the posterior probability  $r_{ni}$  of expert  $i$  producing the label  $y$  for datapoint  $n$ . We will also refer to this as the responsibility of expert  $i$  for datapoint  $n$ . [1 point]
- (d) Take the derivative of the log-likelihood w.r.t. the parameters of each expert  $\boldsymbol{\theta}_i$  and the parameters of the routing mechanism for each expert  $\boldsymbol{\phi}_i$ . Do not substitute expressions for the probabilities but rather provide your answer in terms of  $p(y_n|\mathbf{x}_n, z_n, \boldsymbol{\theta}_i)$ ,  $p(z_n = k|\mathbf{x}_n, \boldsymbol{\Phi})$ . Make sure to express the derivatives in terms of the responsibilities of each expert  $r_{ni}$ . [2 points]

*Hint:  $\frac{\partial f(x)}{\partial x} = f(x) \frac{\partial \log f(x)}{\partial x}$ .*

- (e) Now insert the explicit expression for each of the respective probability distributions and compute the final derivatives for  $\boldsymbol{\theta}_i, \boldsymbol{\phi}_i$ . [2 points]
- (f) Write down an iterative algorithm that maximizes the log-probability of the data by jointly optimizing the  $\boldsymbol{\Theta}$  and  $\boldsymbol{\Phi}$  parameters. Make use of appropriate convergence criteria. [1 point]
- (g) For this question assume that instead of having the prediction for  $\mathbf{x}_n$  be determined by one expert model, we mix the predictions of our experts. Specifically, to compute our final prediction  $\hat{y}_n$  for datapoint  $\mathbf{x}_n$ , we will now weigh the prediction of each expert by its relevancy. Write down the formula for determining the final prediction  $\hat{y}_n$ . Denote the prediction of expert  $k$  for datapoint  $n$  with  $\hat{y}_{nk}$ . [0.5 point]