

1. Suppose that  $\theta$  is an unknown parameter that we need to estimate based on 5 samples, and the samples  $\mathcal{D} = \{X_1, \dots, X_5\}$  are drawn independently according to a fixed conditional probability density function  $p_X(\cdot|\theta)$ , where  $\theta$  is the unknown parameter.

- (a) Assume that, for fixed  $\theta \in (0, \infty)$ , the samples  $X_k, k = 1, \dots, 5$ , have a conditional probability density function given by

$$p_X(x|\theta) = \begin{cases} 1/\theta & \text{if } 0 \leq x \leq \theta \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Let  $x_k$  be the value taken by  $X_k$ ,  $k = 1, \dots, 5$ . Plot the likelihood  $p(\mathcal{D}|\theta)$  in the range  $0 \leq \theta \leq 1$  when  $\max_{k=1,\dots,5} x_k = 0.6$ . Show that the maximum likelihood estimate  $\theta_{ML}$  is  $\max_{k=1,\dots,5} x_k$ , i.e., the largest value among the 5 samples.

- (b) Suppose that the conditional probability density of the samples is given by

$$p_X(x|\theta) = \begin{cases} \theta e^{-\theta x} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Determine the maximum likelihood estimate  $\theta_{ML}$ .

2. The data for this problem can be found in the file ‘Pizza.csv’. The file contains the values of 7 features (moisture (moi), protein (prot), fat, ash, sodium, carb, and calories (cal)) for 300 selected pizzas. Each row (starting with the second row) is a sample that contains the brand of the selected pizza, its ID, and the feature values.

- (a) Compute the mean feature vector  $\mathbf{m}$  (i.e., the vector that contains the average of each feature).
- (b) Compute the scatter matrix and the first two principal components  $\mathbf{w}^1$  and  $\mathbf{w}^2$  with unit length.
- (c) For each sample (or feature vector)  $\mathbf{x}^k, k = 1, \dots, 300$ , find the vector  $\tilde{\mathbf{x}}^k = \mathbf{m} + a_k^1 \mathbf{w}^1 + a_k^2 \mathbf{w}^2$  that minimizes  $\|\mathbf{x}^k - \tilde{\mathbf{x}}^k\|$ . Print out the first two samples and compare to their approximations.
- (d) [Suggested Exercise:] If you compute the sample variance of each feature, where the sample variance for the  $i$ -th feature is given by  $\frac{\sum_{k=1}^{300} (x_{k,i} - m_i)^2}{300-1}$ , some features have much larger variance than others. Oftentimes, we first subtract the mean and then divide by the standard deviation (the square root of variance) before applying PCA. This ensures that all the features are along the same scale. This is called standardizing the data (or standardization). Repeat parts (b) and (c) after standardization.