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

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

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

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

   Ans: The plot of the likelihood is shown i Figure 1 In general, given the samples $x_1, \ldots, x_5$, the likelihood is given by

   $$p(\mathcal{D}|\theta) = \begin{cases} 0 & \text{if } \theta < \max_{k=1,...,5} x_k \\ \frac{1}{\theta^5} & \text{if } \theta \ge \max_{k=1,...,5} x_k \ . \end{cases}$$

   Since θ^{-5} is strictly positive and decreasing in θ over $[\max_{k=1,...,5} x_k, \infty)$, it is clear that the maximum is achieved at $\theta_{ML} = \max_{k=1,...,5} x_k$.

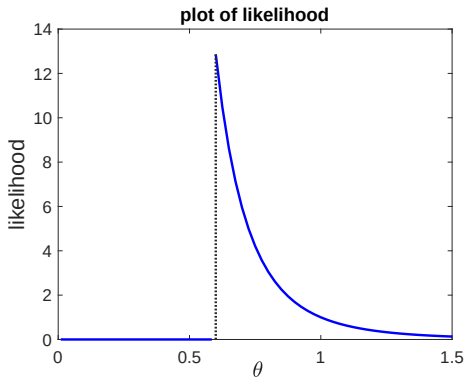

Figure 1: Plot of likelihood for problem 1(a).

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

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

   Determine the maximum likelihood estimate θ_{ML}.

   Ans: In this case, the likelihood is given by

   $$p(\mathcal{D}|\theta) = \theta^5 e^{-\theta \sum_{k=1}^5 x_k} \ .$$

Differentiating it and setting the derivative to 0, we get

$$5\theta^4 e^{-\theta\sum_{k=1}^5 x_k} + \theta^5 e^{-\theta\sum_{k=1}^5 x_k}\Big(-\sum_{k=1}^5 x_k\Big) = \theta^4 e^{-\theta\sum_{k=1}^5 x_k}\Big(5-\theta\sum_{k=1}^5 x_k\Big) = 0\,.$$

The solution must satisfy $5-\theta\sum_{k=1}^5 x_k = 0$ or, equivalently, $\theta_{ML} = 5/\sum_{k=1}^5 x_k$, which is the inverse of the sample mean.

We can do this somewhat more easily if we use the log-likelihood instead.

$$\log(p(\mathcal{D}|\theta) = 5\log(\theta) - \theta\sum_{k=1}^5 x_k$$

If we differentiate it and set the derivative to 0,

$$\frac{5}{\theta} - \sum_{k=1}^5 x_k = 0,$$

which yields the same answer $\theta_{ML} = 5/\sum_{k=1}^5 x_k$.

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.

   Ans: The Matlab code for the parts (a)-(c) is provided on the next page. The mean vector is `meanVec`, the scatter matrix is `ScatterMat`, and the approximated feature vectors are contained in the matrix `featureApprox` as rows.

   (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) $\mathbf{x}^k$, $k = 1, \ldots, 300$, 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.

```
% HW 3 Problem 2, parts (a) - (c)

clear all ;
A = readmatrix('Pizza.csv') ;
featureMat = A(:,3:9) ; % feature vector matrix
sizeFM = size(featureMat) ;
nSample = sizeFM(1) ; % # of samples
nFeature = sizeFM(2) ; % # of features

meanVec = mean(featureMat) ;
featureCent = featureMat - repmat(meanVec, [nSample 1]) ;
ScatterMat = featureCent' * featureCent ;

[V, D] = eig(ScatterMat) ; % V contains eigenvectors, diag(D) contains eigenvalues
disp('Eigenvalues of scatter matrix')
diag(D)'
PC = zeros(nFeature, 2) ; % will contain the first two PCs
PC(:,1) = V(:, nFeature) / norm(V(:, nFeature)) ;  % first normalized PC
PC(:,2) = V(:, nFeature - 1) / norm(V(:, nFeature-1));  % second normalized PC

% Construct the approximations using the two PCs

weights = featureCent * PC ; % coefficients for projections

featureApprox = weights * PC' + repmat(meanVec, [nSample 1]) ;

% Compare the first two samples
featureMat(1:2, :)
featureApprox(1:2, :)
```