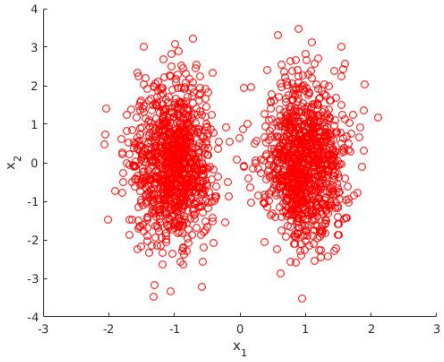

Figure 1: Two-dimensional data set.

1. Describe a step-by-step procedure that calculates a PCA transformation matrix $\boldsymbol{P}$ of the n-dimensional sample $\boldsymbol{X}$. This transformation $\boldsymbol{P}$ should retain as many principal components k as necessary in order to push the expected reconstruction error below a certain threshold η (e.g., 10%).

2. You are given a data set $\boldsymbol{X} \in \mathbb{R}^2$ with mean $\boldsymbol{m} = [1, 4]^\top$ and covariance matrix $\boldsymbol{S} = \begin{bmatrix} 5 & -3 \\ -3 & 5 \end{bmatrix}$. Furthermore, you are given a point $\boldsymbol{x} = [5, 3]^\top$ Compute:

   - The projection of $\boldsymbol{x}$ onto the one-dimensional principal subspace, such that the expected projection error is minimized
   - The expected projection error
   - The projection error of $\boldsymbol{x}$
   - The coordinates of $\boldsymbol{x}$ with respect to the basis of the principal subspace.

3. Why does mean normalization make sense?

4. What is a problem of PCA if the individual variables possess very different units, e.g., x_1 is measured in m, whereas x_2 is measured in mm? How could you address this problem?

5. The mean and covariance of the data set in Fig. 1 are

$$\boldsymbol{m} = \begin{bmatrix} -0.0181 \\ -0.0045 \end{bmatrix}, \qquad \boldsymbol{S} = \begin{bmatrix} 1.0958 & 0.0088 \\ 0.0088 & 1.0956 \end{bmatrix}, \tag{1}$$

   respectively.

   What is (approximately) the smallest (average) reconstruction error if we project the data onto a one-dimensional subspace (no computations required)? Briefly justify your answer.

6. The data in Fig. 2 was generated from a mixture model. A central problem with mixture models is to determine the number of mixture components. Briefly describe an

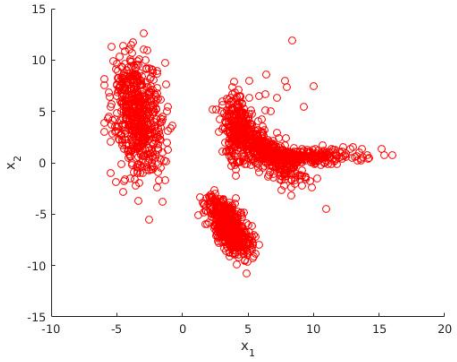

Figure 2: Data generated from a mixture model

algorithm for choosing the number of mixture components. Briefly discuss one strength and one weakness of this algorithm.