Assignment 4

Problem 1

Learn to generate synthetic data with Gaussian probability distribution specified by user entered parameters

Generate 1000 samples each with 2 dimensions, say this data matrix $\mathbf{D}$. Each sample is independently and identically distributed with multi-variate (multi ≥ 2 dimensions) Gaussian distribution with user entered mean values and covariance matrix ($\boldsymbol{\mu} = [\mu_1, \mu_2], \boldsymbol{\Sigma} \in \mathbb{R}^{2\times 2}$).

Draw a scatter plot of the data samples. Observe and relate the distribution of data samples in the plot and parameters of distribution.

1. Relate the data plot with entered mean, variances and co-variances terms.
2. What do you expect regarding the directions of Eigenvectors for this data?
3. Assuming each value in data consumes p bytes, how many bytes are required for 1000 samples consisting of 2 dimensions? Save the data matrix $\mathbf{D}$ and note its memory usage.

*Hint: Use **np.random.multivariate_normal** for generating the multivariate normal distribution.*

Problem 2

Dimension Reduction - An initial perspective

Compute a matrix $\mathbf{A} \in \mathbb{R}^{2\times 1}$ such that $\mathbf{D} \times \mathbf{A} = \hat{\mathbf{D}}$. Here $\mathbf{D}$ is the data matrix with data samples in its columns (2 dimensions) and $\hat{\mathbf{D}}$ represents the matrix with reduced dimensions (1 dimension). How can you generate such matrix $\mathbf{A}$? Compute the reconstruction error between $\hat{\mathbf{D}}$ obtained using (your generated) $\mathbf{A}$, and $\mathbf{D}$ using mean square error.

Hint: Think of $\mathbf{A}$ as $\mathbf{A} = [a_1\ a_2]$, and a_1 & a_2 as the weights to the first and second column of matrix $\mathbf{D}$ respectively, while multiplying $\mathbf{D}$ with $\mathbf{A}$.

Problem 3

Compute the Eigenvectors and show it on the scatter plot.

Compute co-variance matrix and then compute eigenvectors and eigenvalues.

1. Reconstruct the data samples using all eigenvectors, say it $\hat{\mathbf{D}}$. Compute the reconstruction error between $\hat{\mathbf{D}}$ and $\mathbf{D}$ using mean square error.

Assignment 4

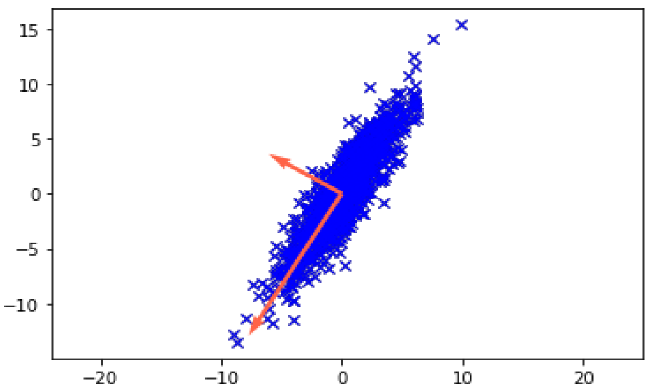


Figure 1: Blue points denotes the data samples and red arrow shows the eigen directions. The parameters for shown data is $\boldsymbol{\mu} = [0\ 0]$ and $\boldsymbol{\sigma} = [[7, 10], [10, 18]]$

2. Draw the scatter plot of data samples. Plot the eigen directions (with arrows/lines) onto the scatter plot of data, as shown in Figure. 1.
3. Observe the directions of eigenvectors. Check the eigenvalues and see if is similar to variance values of projected data.
4. Observe the covariance matrix of the projected data and write down your inferences.
5. Which eigen direction shall we omit, while reconstring $\hat{\mathbf{D}}$, in order to reduce the dimensionality of data?

- *Use **numpy.linalg.eig** function to compute the eigen vectors.*
- *Use matplot **quiver** function for plotting arrows in eigen directions.*

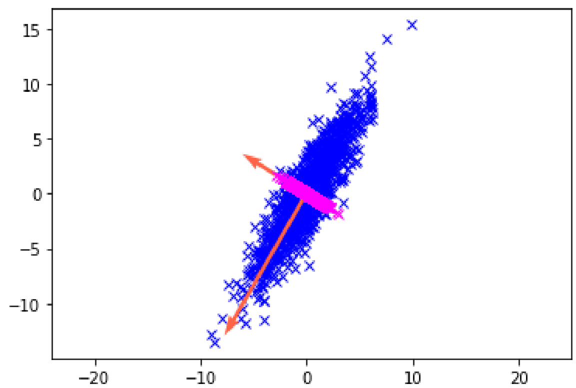
Problem 4

Reduce the dimensions.

Reconstruct the data samples using only one eigenvectors instead of two. This will reduce the dimension of the data matrix.

1. Project the data onto first and second eigen direction one by one, say it $\hat{\mathbf{D}}_1$ and $\hat{\mathbf{D}}_2$ respectively. Draw the scatter plot of $\hat{\mathbf{D}}_1$ and $\hat{\mathbf{D}}_2$ separately onto figure vreated in Problem 3 - Part 2, as shown in Figure. 2.
2. Compute the number of bytes required to save this reduced dimension matrix. Compare it with memory required for data matrix A. Save the data matrix and note its memory usage.

Assignment 4



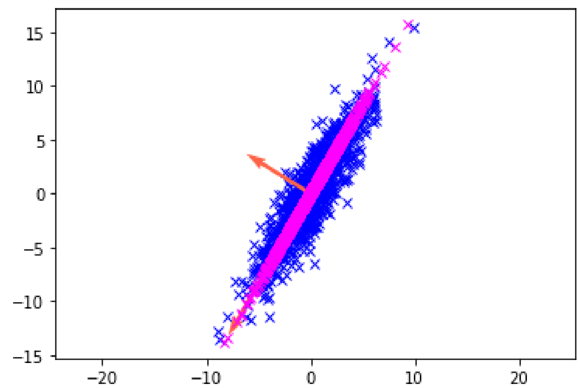


Figure 2: Pink points show the projected values onto the first and second eigen directions in left and right images respectively.

3. Observe the changes in behavior of data distribution.
4. Compute the reconstruction error for $\hat{\mathbf{D}}_1$ and $\hat{\mathbf{D}}_2$, using mean square error and compare it with the error obtained in problem 2.
5. What is your $\mathbf{A}$ here ($\mathbf{A}$ defined in problem 2). Compare the significance of both.

To Think

- What do eigenvectors and eigenvalues represent?
- Do the variance values of projected data and eigen values match?
- Which eigenvectors are least significant for defining the data distribution?
- Suggest the cases in which PCA based dimensionality reduction is not good.