PCA Quiz Questions

**Question 1.** We have a data set $\mathcal{D} = \{(1, 2)^T, (7, 3)^T, (3, 4)^T, (-1, -1)^T\} \subset \mathbb{R}^2$. Let $\mathbf{X} = (X_1, X_2)^T$ be a r.v. with $X_1$ and $X_2$ random variables for the $x_1$ and $x_2$ co-ordinates respectively.

i) Find $\hat{\mathbb{E}}[\mathbf{X}]$.

ii) Find $\hat{\text{Var}}[X_1], \hat{\text{Var}}[X_2]$.

iii) Find $\hat{\text{Cov}}[X_1, X_2]$.

*(give your answers in 3 significant figures)*

**Question 2.** We have a data set $\mathcal{D} = \{(-3, -7, -1)^T, (2, -1, -6)^T, (6, -2, 6)^T, (-5, 16, -1)^T, (-3, -6, 4)^T, (2, -7, 2)^T, (4, -1, -2)^T, (-3, 8, -2)^T\} \subset \mathbb{R}^3$. Let $\mathbf{X} = (X_1, X_2, X_3)^T$ be a r.v. Find the co-variance matrix $\text{Cov}[\mathbf{X}]$ *(give your answers in 3 significant figures)*.

**Question 3.** Suppose that we have a 2-dimensional data set $\mathcal{D}$ and we plot the points on a Cartesian graph, producing the following shape as shown:

![Cartesian Graph](https://via.placeholder.com/150)
Let $X = (X_1, X_2)^T$ be a r.v. What do you expect the co-variance matrix $\text{Cov}[X]$ to look like?

**Question 4.** Suppose we have the $3 \times 3$ square matrix

$$
A = \begin{pmatrix}
1 & 4 & 5 \\
3 & 2 & 2 \\
4 & 2 & 3 
\end{pmatrix}
$$

Find all distinct eigenvectors of $A$ and their respective eigenvalues (give your answers in 3 significant figures).

**Question 5.** Let $D$ be a 2-dimensional data set of size $N$ and $X = (X_1, X_2)^T$ be a r.v, with $X_1$ and $X_2$ random variables for the $x_1$ and $x_2$ co-ordinates respectively. Suppose that for each $X_i$ we have calculated the mean $\mu_i$ and variance $\sigma_i$ using $D$. For $i = 1, 2$, Let $\tilde{X}_i$ be a new r.v given by

$$
\tilde{X}_i = \frac{X_i - \mu_i}{\sigma_i}.
$$

Show that $\text{Cov}[\tilde{X}_1, \tilde{X}_2] = \text{Corr}(X_1, X_2)$, where $\text{Corr}(X_1, X_2)$ is the sample correlation coefficient between $X_1$ and $X_2$ defined as

$$
\text{Corr}(X_1, X_2) = \frac{\text{Cov}[X_1, X_2]}{\sigma_1 \sigma_2}.
$$

**Question 6.** Suppose that we perform PCA on a $d$-dimensional data set $D$, computing the eigenvectors $v_1, v_2, \ldots, v_d$ and their respective eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ in the process.

i) When projecting $D$ onto a lower-dimensional data set, explain what it means when we want to preserve a proportion $\rho$ of the variability in the data.

ii) Suppose that $d = 4$ and we find the following eigenvectors $v_1, \ldots, v_4$ and their corresponding eigenvalues:

$$
\lambda_1 = 2.97, \\
\lambda_2 = 1.17, \\
\lambda_3 = 0.64, \\
\lambda_4 = 0.2.
$$

Which eigenvectors should be picked for our new axes so that 90% of the variation in $D$ is preserved?
**Question 7.** In this question we will apply the PCA algorithm discussed in lectures using a 2-dimensional data set.

Let $\mathcal{D}$ be a data set given by

$$\mathcal{D} = \{(-2,-3), (0,-2), (1,-1.5), (-1,-0.5), (1,-0.5),$$
$$ (0,-0.5), (4,-2), (1,0), (3,0), (3,1), (1,3),$$
$$ (3.5,3), (4,3), (5,1)\}.$$

Let $\mathbf{X} = (X_1, X_2)^T$ be a r.v.

i) Find the co-variance matrix $\mathbf{C}$.

ii) Find the normalised eigenvectors of $\mathbf{C}$ and their corresponding eigenvalues. Write $\mathbf{C}$ in the form $\mathbf{V} \cdot \tilde{\mathbf{C}} \cdot \mathbf{V}^T$, where $\mathbf{V}$ and $\tilde{\mathbf{C}}$ are matrices as described in the lectures.

iii) Choose the eigenvector that will preserve the most variation in $\mathcal{D}$ and project the data points onto the new axis.

(give your answers in 3 significant figures)

**Question 8.** Let $\mathbf{C}$ be the co-variance matrix of a $d$-dimensional r.v $\mathbf{X} = (X_1, X_2, \ldots, X_d)^T$.

i) Show that $\mathbf{C} = \mathbb{E}(\mathbf{XX}^T) - \mathbf{\mu}\mathbf{\mu}^T$, where $\mathbf{\mu} = \mathbb{E}(\mathbf{X})$.

ii) Prove that the eigenvalues of $\mathbf{C}$ will always be positive.