Graded

This document has 11 questions.

Note to Learners

Statement

The projection is treated as a vector in all the questions. If we wish to talk about the length of the projection, then that would be mentioned explicitly. Likewise, the residue after the projection is also treated as a vector. If we wish to talk about the length of the residue, that would be mentioned explicitly.

Question-1 [0.5 point]

Statement

Consider a point $\mathbf{x} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ and a line passing through the origin which is represented by the vector $\mathbf{w} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$. What can you say about the following quantities? (MSQ)

(1) the projection of $\mathbf{x}$ onto the line

(2) the residue

Options

(a)

The residue is equal to the zero vector.

(b)

The residue is equal to the vector $\mathbf{x}$.

(c)

The projection is the zero vector.

(d)

The projection is equal to the vector $\mathbf{x}$.

Answer

(b), (c)

Solution

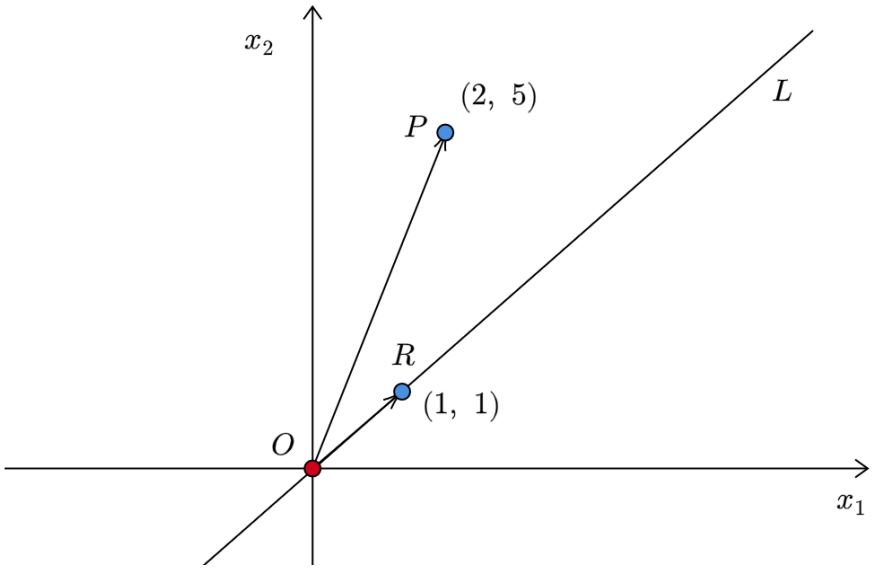
We have $\mathbf{x}^T\mathbf{w} = 0$. So, the projection is the zero vector. The residue is given by:

$$\mathbf{x} - (\mathbf{x}^T\mathbf{w})\mathbf{w} = \mathbf{x}$$

Common data for questions (2) to (5)

Statement

Consider a point P and a line L that passes through the origin O. The point R lies on the line.



We use the following notation:

$$\mathbf{w} = \overrightarrow{OR}$$

$$\mathbf{x} = \overrightarrow{OP}$$

Question-2 [1 point]

Statement

Consider the following statements:

Statement-1: The projection of $\mathbf{x}$ on the line L is given by $(\mathbf{x}^T\mathbf{w})\mathbf{w}$

Statement-2: The projection of $\mathbf{x}$ on the line L is given by $(\mathbf{x}^T\mathbf{w})\mathbf{x}$

Statement-3: The projection of $\mathbf{x}$ on the line L is given by $(\mathbf{x}^T\mathbf{x})\mathbf{w}$

Statement-4: The projection of $\mathbf{x}$ on the line L is given by $\mathbf{w}^T\mathbf{x}$

Which of the above statements is true?

Options

(a)

Statement-1

(b)

Statement-2

(c)

Statement-3

(d)

Statement-4

(e)

None of these statements are true.

Answer

(e)

Solution

The projection of a point $\mathbf{x}$ on a line $\mathbf{w}$ is given by:

$$\frac{\mathbf{x}^T\mathbf{w}}{\mathbf{w}^T\mathbf{w}}\mathbf{w}$$

This is the expression when $\mathbf{w}$ does not have unit length. In this problem, $\mathbf{w}$ does not have unit length. If $||\mathbf{w}|| = 1$, then the expression becomes:

$$(\mathbf{x}^T\mathbf{w})\mathbf{w}$$

Question-3 [1 point]

Statement

Find the length of the projection of $\mathbf{x}$ on the line L. Enter your answer correct to two decimal places.

Answer

4.95

Range: $[4.9, 5.0]$

Solution

The length of the projection is given by:

$$\frac{|\mathbf{x}^T\mathbf{w}|}{||\mathbf{w}||} = \frac{2+5}{\sqrt{2}} \approx 4.95$$

Question-4 [1 point]

Statement

Find the residue after projecting $\mathbf{x}$ on the line L.

Options

(a)

$$\begin{bmatrix} 3.5 \\ 3.5 \end{bmatrix}$$

(b)

$$\begin{bmatrix} -1.5 \\ 1.5 \end{bmatrix}$$

(c)

$$\begin{bmatrix} 2 \\ 5 \end{bmatrix}$$

(d)

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Answer

(b)

Solution

The residue is given by:

$$\mathbf{x} - \frac{\mathbf{x}^T\mathbf{w}}{\mathbf{w}^T\mathbf{w}}\mathbf{w} = \begin{bmatrix} 2 \\ 5 \end{bmatrix} - \frac{7}{2} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1.5 \\ 1.5 \end{bmatrix}$$

Question-5 [1 point]

Statement

Find the reconstruction error for this point. Enter your answer correct to two decimal places.

Answer

4.5

Range: $[4.4, 4.6]$

Solution

The reconstruction error is given by the square of the length of the residue. If the residue is $\mathbf{x}'$, then:

$$(\mathbf{x}')^T\mathbf{x}' = (-1.5)^2 + 1.5^2 = 4.5$$

Programming based solution. This is to be used only to verify the correctness of the calculations. The added benefit is that you get used to `NumPy`.

```python
import numpy as np

x = np.array([2, 5])
w = np.array([1, 1])
w = w / np.linalg.norm(w)

# Projection
proj = (x @ w) * w
print(f'Projection = {np.linalg.norm(proj)}')
# Residue
res = x - proj
print(f'Residue = {res}')
# Reconstruction error
recon = res @ res
print(f'Reconstruction error = {recon}')
```

Question-6 [0.5 point]

Statement

Consider the following images of points in 2D space. The red line segments in one of the images represent the lengths of the residues after projecting the points on the line L. Which image is it?

Image-1

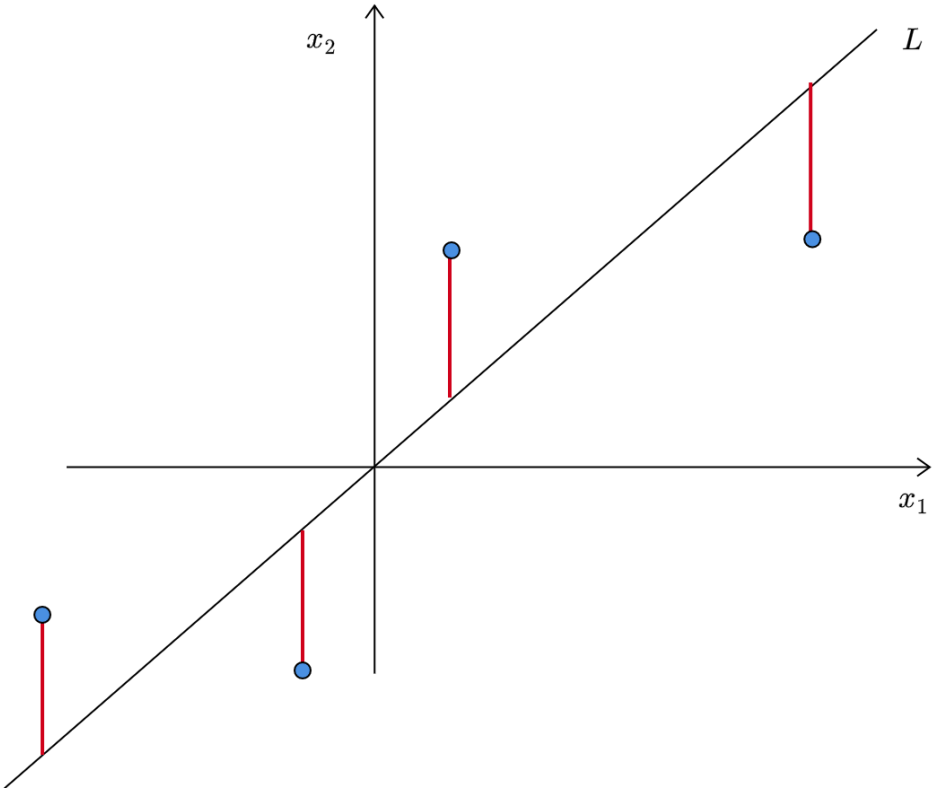
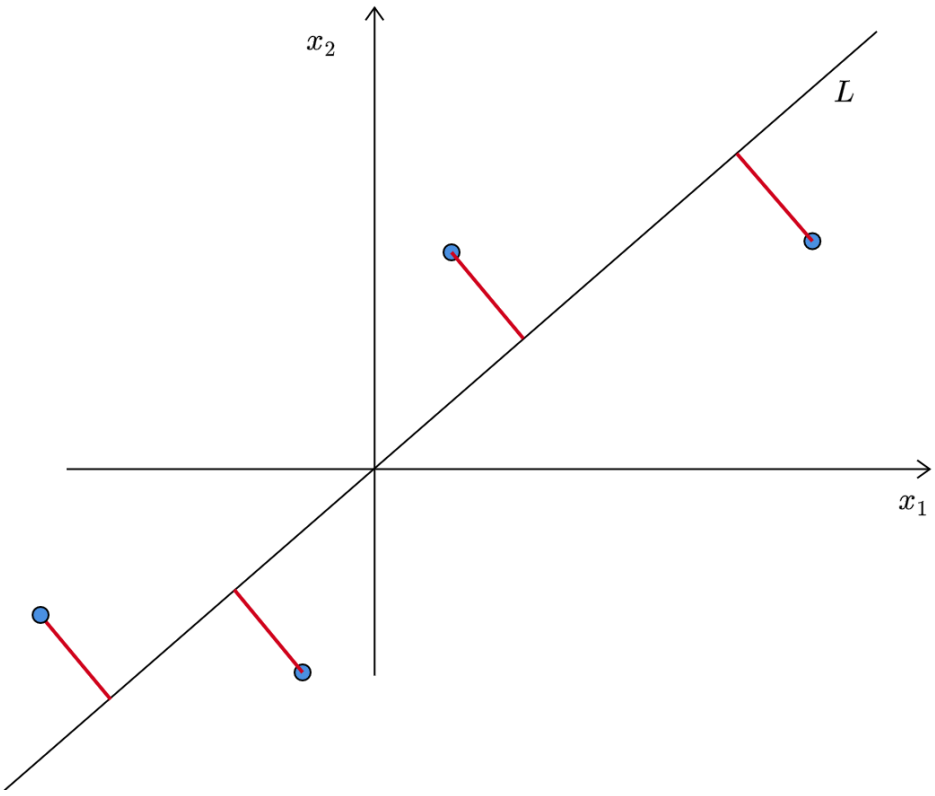


Image-2



Options

(a)

Image-1

(b)

Image-2

Answer

(b)

Solution

The residue after the projection should be perpendicular to the line. Note that by projection we mean the orthogonal projection of a point on a line. The projection of a point on a line is one of the proxies for that point on the line, in fact it is the "best" possible proxy. But every proxy does not become a projection. The projection of a point on a line is unique.

Question-7 [1 point]

Statement

Consider a dataset that has 1000 samples, where each sample belongs to $\mathbb{R}^{30}$. PCA is run on this dataset and the top 4 principal components are retained, the rest being discarded. If it takes one unit of memory to store a real number, find the percentage decrease in storage space of the dataset by moving to its compressed representation. Enter your answer correct to two decimal places; it should lie in the range $[0, 100]$.

Answer

86.27

Range: $[86.2, 86.3]$

Solution

Original space = $1000 \times 30 = 30000$

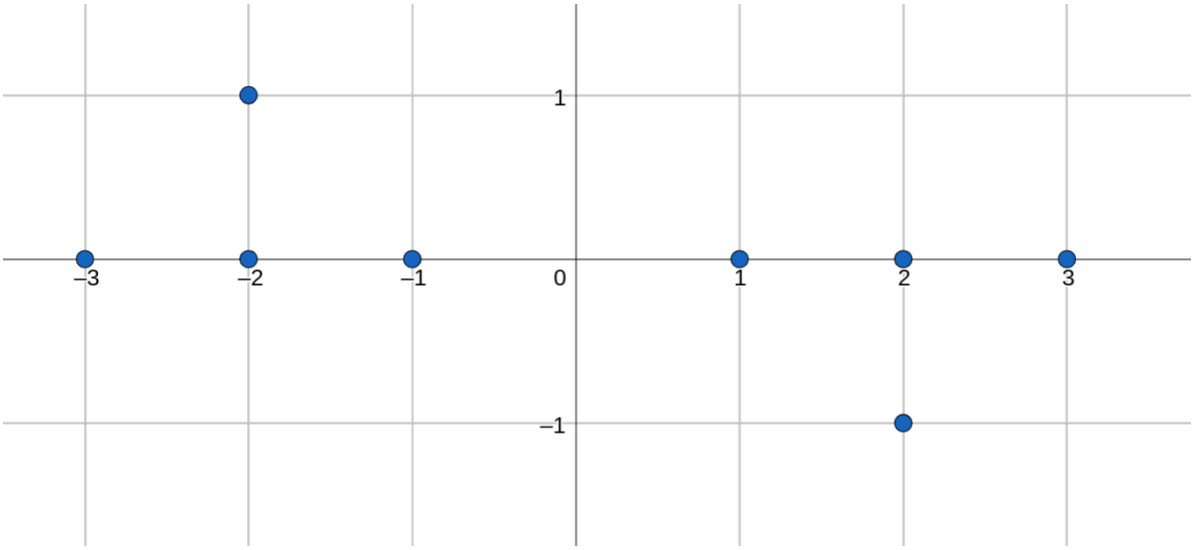
Compressed space = $1000 \times 4 + 4 \times 30 = 4120$

Percentage decrease in space = $\frac{30000 - 4120}{30000} \times 100 \approx 86.27$

Common Data for questions (8) to (9)

Statement

Consider a dataset that has 8 points all of which belong to $\mathbb{R}^2$:



Question-8 [1 point]

Statement

Find the covariance matrix of this dataset.

Options

(a)

$$\begin{bmatrix} 4.5 & -0.5 \\ -0.5 & 0.25 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 36 & -4 \\ -4 & 2 \end{bmatrix}$$

(c)

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

(d)

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Answer

(a)

Solution

Let us first arrange the data in the form of a $n \times d$ matrix. Here, $n = 8$ and $d = 2$:

$$\mathbf{X} = \begin{bmatrix} -3 & 0 \\ -2 & 0 \\ -2 & 1 \\ -1 & 0 \\ 1 & 0 \\ 2 & 0 \\ 2 & -1 \\ 3 & 0 \end{bmatrix}$$

The covariance matrix is therefore:

$$\frac{1}{n} \cdot \mathbf{X}^T \mathbf{X} = \begin{bmatrix} 4.5 & -0.5 \\ -0.5 & 0.25 \end{bmatrix}$$

Question-9 [1 point]

Statement

If PCA is run on this dataset, find the variance of the dataset along the first principal component. The eigenvectors of the covariance matrix are given below:

$$\begin{bmatrix} -0.993 \\ 0.115 \end{bmatrix}, \quad \begin{bmatrix} -0.115 \\ -0.993 \end{bmatrix}$$

Recall that the first principal component is the most important. Enter your answer correct to two decimal places.

Answer

4.55

Range: $(4.5, 4.6)$

Solution

If $(\lambda_k, \mathbf{w}_k)$ is the k^{th} eigenpair for $\mathbf{C}$, we have:

$$\lambda_k = \mathbf{w}_k^T \mathbf{C} \mathbf{w}_k$$

Of the two eigenvalues, the larger one is the answer.

Verification for these two problems:

```
import numpy as np

X = np.array([[-3, 0],
              [-2, 0],
              [-2, 1],
              [-1, 0],
              [1, 0],
              [2, 0],
              [2, -1],
              [3, 0]])

C = X.T @ X / X.shape[0]
print(f'Covariance matrix = {C}')
eigval, eigvec = np.linalg.eigh(C)
print(f'Variance = {eigval[-1]}')
```

A more detailed version. The variance of the dataset along the j^{th} principal component is σ_j^2 and is given by:

$$\sigma_j^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (\mathbf{x}_i^T \mathbf{w}_j)^2$$

$$= \mathbf{w}_j^T \left(\frac{1}{n} \cdot \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{w}_j$$

$$= \mathbf{w}_j^T \mathbf{C} \mathbf{w}_j$$

$$= \lambda_j$$

So, the variance along the j^{th} principal component is the j^{th} largest eigenvalue of the covariance matrix.

Question-10 [1 point]

Statement

Consider a dataset of 100 points all of which lie in $\mathbb{R}^5$. The eigenvalues of the covariance matrix are given below:

$$3.4, \quad 2.8, \quad 0.5, \quad 0.4, \quad 0.01$$

If we run the PCA algorithm on this dataset and retain the top-k principal components, what is a good choice of k? Use the heuristic that was discussed in the lectures.

Answer

4

Solution

The top-k principal components should capture 95% of the variance. Here is a code snippet to answer this question:

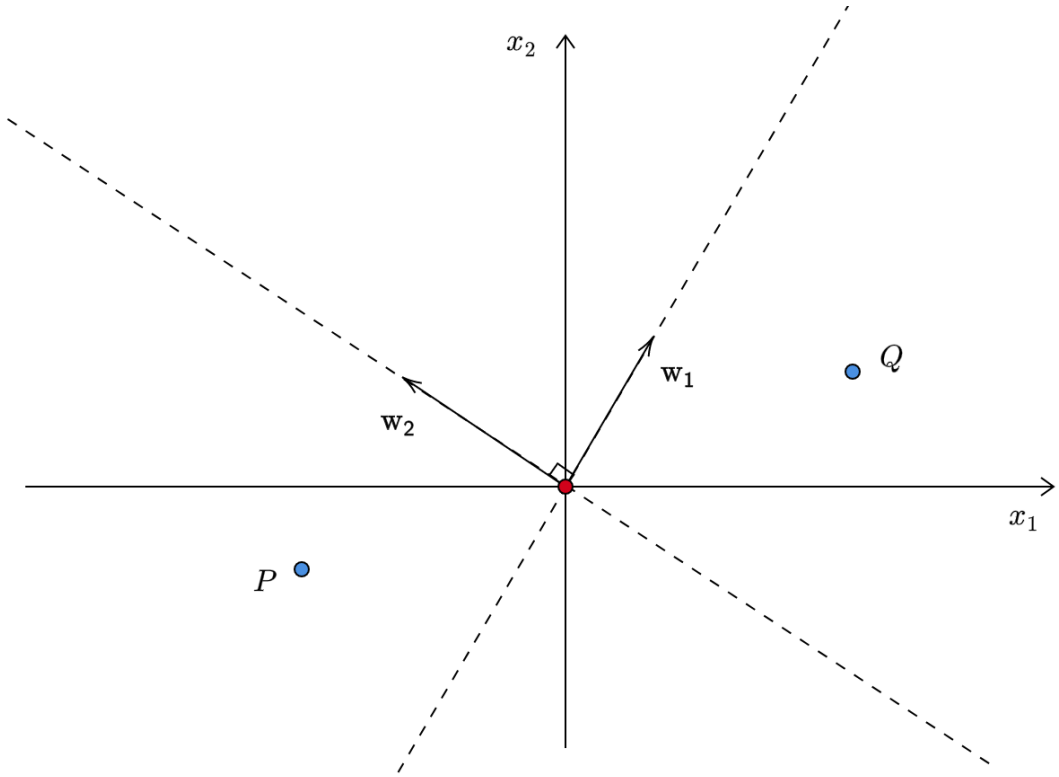
```
L = [3.4, 2.8, 0.5, 0.4, 0.01]

den = sum(L)
for k in range(1, len(L) + 1):
    num = sum(L[: k])
    if num / den >= 0.95:
        break
print(k)
```

Question-11 [1 point]

Statement

PCA is run on a dataset that has 2 features. The resulting principal components are $\mathbf{w}_1$ and $\mathbf{w}_2$. We represent the points in 2D space in terms of this new coordinate system made up of the principal components. The first coordinate corresponds to $\mathbf{w}_1$ and the second to $\mathbf{w}_2$. In such a scenario, what would be the sign of the coordinates for the points P and Q?



Options

(a)

$P : (-ve, -ve)$

(b)

$P : (-ve, +ve)$

(c)

$Q : (+ve, +ve)$

(d)

$Q : (+ve, -ve)$

Answer

(b), (d)

Solution

Each vector $\mathbf{w}$ is associated with a line perpendicular to it. This line divides the space into two halves. The basic idea is to identify the sign of the half-planes into which the line perpendicular to the vector $\mathbf{w}$ divides the space.