Problem 1 (30%). Consider the dataset as shown in the figure below. What is the covariance matrix of the dataset?

Answer: Let \( A = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} \) be the covariance matrix, where \( \sigma_{xx} \) (\( \sigma_{yy} \)) is the variance along the x- (y-) dimension, and \( \sigma_{xy} \) (\( = \sigma_{yx} \)) is the covariance of the x- and y-dimensions. Since the means along both the x- and y-dimensions are 0, we have that:

\[
\sigma_{xx} = \frac{1}{4}((-3)^2 + (-2)^2 + 1^2 + 4^2) = 30/4 = 7.5
\]
\[
\sigma_{yy} = \frac{1}{4}(4^2 + 1^2 + (-2)^2 + (-3)^2) = 30/4 = 7.5
\]
\[
\sigma_{xy} = \frac{1}{4}((-3) \times 4 + (-2) \times 1 + 1 \times (-2) + 4 \times (-3)) = -28/4 = -7
\]

Therefore, \( A = \begin{bmatrix} 7.5 & -7 \\ -7 & 7.5 \end{bmatrix} \).

Problem 2 (30%). Use PCA to find the line passing the origin on which the projections of the points in Problem 1 have the greatest variance.

Answer: Let \( \lambda \) be an eigenvalue of \( A \), which implies that the determinant of 

\[
\begin{vmatrix} 7.5 - \lambda & -7 \\ -7 & 7.5 - \lambda \end{vmatrix}
\]

is 0. By expanding the determinant, we get the following equation:

\[
(7.5 - \lambda)^2 - 49 = 0.
\]

It follows that \( \lambda_1 = 14.5 \) and \( \lambda_2 = 0.5 \) are the eigenvalues of \( A \), where \( \lambda_1 \) is the larger one.

Let \( \vec{v} = \begin{bmatrix} x \\ y \end{bmatrix} \) be an eigenvector corresponding to \( \lambda_1 \), which satisfies that

\[
\begin{bmatrix} 7.5 & -7 \\ -7 & 7.5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 14.5x \\ 14.5y \end{bmatrix}.
\]
Note that the above equation is satisfied by any pair of \( x \) and \( y \) satisfying \( x + y = 0 \). As the line chosen by PCA has the same direction as \( \vec{v} \), the line is \( x + y = 0 \).

**Problem 3 (40%)**. Run DBSCAN on the set of points shown in the figure below with \( \epsilon = 1 \) and \( \text{minpts} = 4 \). What are the core points and the clusters?

![Diagram of a 2D grid with points marked](image)

**Answer**: The core points are \( b, e, g, j, k \) and \( o \). There are three clusters:

- Cluster 1: \( a, b, c, d, e, f \)
- Cluster 2: \( f, g, h, i, j, k, l \)
- Cluster 3: \( m, n, o, p, q \)