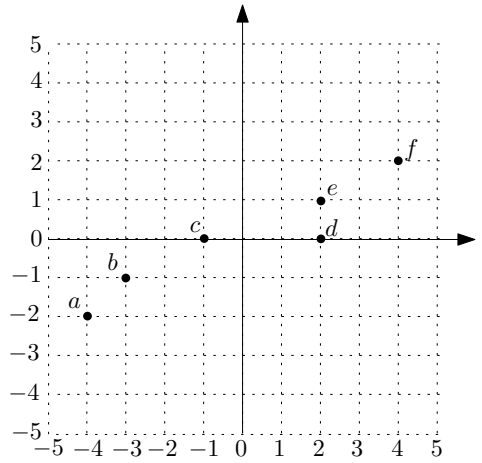
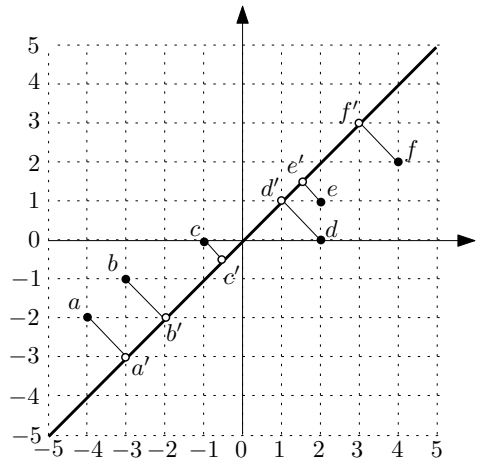
CMSC5724: Exercise List 10

Problem 1. Consider the following 2D dataset:



What is the variance of the projections of the 6 points onto the line $y = x$?

Answer. The projections are as shown below:



The projection of, for example, a is a', which we will regard as a 1D value, equal to the negated distance $-3\sqrt{2}$ from a' to the origin (it is negated because it is below the y-axis). Similarly, the projections b', c', d', e' and f' can be regarded as 1D values: $-2\sqrt{2}, -\sqrt{2}/2, \sqrt{2}, \frac{3}{2}\sqrt{2}$ and $3\sqrt{2}$, respectively. The mean of the six 1D values is 0; hence, their variance is:

$$\begin{aligned} & \frac{1}{6}((-3\sqrt{2}-0)^2 + (-2\sqrt{2}-0)^2 + (-\sqrt{2}/2-0)^2 + (\sqrt{2}-0)^2 + (\frac{3}{2}\sqrt{2}-0)^2 + (3\sqrt{2}-0)^2) \\ = \; & (18+8+0.5+2+4.5+18)/6 \\ = \; & 51/6 \end{aligned}$$

Problem 2. What is the co-variance matrix of the dataset in Problem 1?

Answer. The co-variance matrix is $A = \left[\begin{smallmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{smallmatrix}\right]$, where σ_{xx} (σ_{yy}) is the variance along the x- (y-) dimension, and σ_{xy} ($= \sigma_{yx}$) is the covariance of the x- and y-dimensions. We calculate:

$$\begin{aligned}
\sigma_{xx} &= ((-4-0)^2 + (-3-0)^2 + (-1-0)^2 + (2-0)^2 + (2-0)^2 + (4-0)^2)/6 \\
&= (16+9+1+4+4+16)/6 = 25/3 \\
\sigma_{yy} &= ((-2-0)^2 + (-1-0)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (2-0)^2)/6 \\
&= (4+1+0+0+1+4)/6 = 5/3 \\
\sigma_{xy} &= ((-4)(-2) + (-3)(-1) + (-1)0 + 2 \cdot 0 + 2 \cdot 1 + 4 \cdot 2)/6 \\
&= (8+3+0+0+2+8)/6 = 3.5
\end{aligned}$$

Hence, $A = \left[\begin{smallmatrix} 25/3 & 3.5 \\ 3.5 & 5/3 \end{smallmatrix}\right]$.

Problem 3. Use PCA to find the line passing the origin on which the projections of the points in Problem 1 have the greatest variance.

Answer. We first compute the eigenvectors v of the covariance matrix A. Specifically, there should be a non-zero value λ such that:

$$\begin{aligned}
Av &= \lambda v \quad \Leftrightarrow \\
\begin{bmatrix} 25/3 & 3.5 \\ 3.5 & 5/3 \end{bmatrix} v &= \lambda v \quad \Leftrightarrow \\
\begin{bmatrix} 25/3-\lambda & 3.5 \\ 3.5 & 5/3-\lambda \end{bmatrix} v &= 0
\end{aligned}$$

Setting the determinant of $\left[\begin{smallmatrix} 25/3-\lambda & 3.5 \\ 3.5 & 5/3-\lambda \end{smallmatrix}\right]$ to 0 gives:

$$(25/3-\lambda)(5/3-\lambda) = 3.5^2$$

which has two solutions (a.k.a. eigenvalues):

$$\begin{aligned}
\lambda_1 &= 59/6 \\
\lambda_2 &= 1/6
\end{aligned}$$

The line that we are looking for (i.e., the one capturing the most variance of the projections) is given by an eigenvector $v = \left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$ corresponding to the larger eigenvalue λ_1. To find v, we fit $\lambda_1 = 59/6$ into Equation 1 (after simplification):

$$\begin{bmatrix} -1.5 & 3.5 \\ 3.5 & -49/6 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0$$

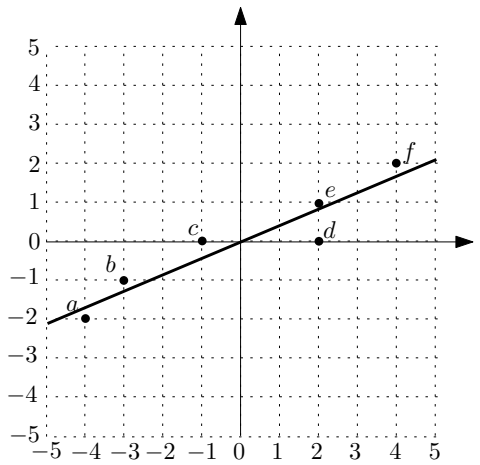
namely:

$$-1.5x + 3.5y = 0$$

It suffices to find any solution of x and y, e.g.:

$$\begin{aligned}
x &= 3.5 \\
y &= 1.5
\end{aligned}$$

The line we are looking for is therefore the one passing origin and point (3.5, 1.5), as shown below:



Remark. In the lecture notes, we required that eigenvectors should be normalized to have length 1. That was for the purpose of proving the correctness. This is not necessary for implementation. In the above, if you want, you can find a unit eigenvector by normalizing the vector $(3.5, 1.5)$ to $(3.5/\sqrt{14.5}, 1.5/\sqrt{14.5})$, but this will not change the the direction of the line.

Problem 4. In the previous problem, suppose that we perform dimensionality reduction using only 1 principle component (i.e., using only the first eigenvector), show the 6 reconstructed points.

Answer. The reconstructed points are the 6 white points in the figure below (the original black points are shown for your convenience).