1 PCA Handout

Given random variables $X$ and $Y$, the covariance between $X$ and $Y$ is denoted as $E[(X - E[X])(Y - E[Y])]$.

To get an empirical estimate of this covariance given samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ we shall use the estimate:

$$\frac{1}{n} \sum_{t=1}^{n} \left( X_t - \frac{1}{n} \sum_{s=1}^{n} X_s \right) \left( Y_t - \frac{1}{n} \sum_{s=1}^{n} Y_s \right)$$

Now given our $d$ dimensional data represented as $d$ dimensional vectors $x_1, \ldots, x_n$, the empirical covariance matrix which we shall denote by the matrix $\Sigma$ is basically the matrix whose $i, j$'th entry is the empirical covariance between the $i$'th and $j$'th coordinates of the data.

Denote the mean of these vectors by $\mu = \frac{1}{n} \sum_{t=1}^{n} x_t$. Show that $\Sigma$ can be written as an average of outer products of vectors as:

$$\Sigma = \sum_{t=1}^{n} (x_t - \mu)(x_t - \mu)\top$$
Let \( w \) be a \( d \) dimensional projection vector and the 1 dimensional projection of points \( x_1, \ldots, x_n \) is obtained by setting

\[ y_t = w^\top x_t \]

If our goal is to find a \( w \) such that \( w \) is unit length (i.e. \( \|w\|_2 = 1 \)) and spread or variance of the \( y \)'s is maximized then show that the optimization problem we need to solve is:

\[
\text{Maximize} \quad w^\top \Sigma w \quad \text{subject to} \quad \|w\|_2 = 1
\]

Start here: We need to find \( w \) s.t. \( \|w\|_2 = 1 \) and it maximizes the spread/variance of \( y \)'s given by:

\[
\text{Variance}(y_1, \ldots, y_n) = \frac{1}{n} \sum_{t=1}^{n} \left( y_t - \frac{1}{n} \sum_{t=1}^{n} y_t \right)^2
\]