1 PCA Handout

Given random variables X and Y, the covariance between X and Y is denoted as $\mathbb{E}\left[(X - \mathbb{E}[X])\,(Y - \mathbb{E}[Y])\right]$. 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 $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the empirical covariance matrix which we shall denote by the matrix Σ 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 $\mu = \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t$. Show that Σ can be written as an average of outer products of vectors as:

$$\Sigma = \sum_{t=1}^{n} (\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^{\top}$$

Let $\mathbf{w}$ be a d dimensional projection vector and the 1 dimensional projection of points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is obtained by setting

$$y_t = \mathbf{w}^\top \mathbf{x}_t$$

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

$$\text{Maximize} \quad \mathbf{w}^\top \Sigma \mathbf{w} \qquad \textbf{subject to } \|\mathbf{w}\|_2 = 1$$

Start here: We need to find $\mathbf{w}$ s.t. $\|\mathbf{w}\|_2 = 1$ and it maximizes the spread/varinace of y's given by:

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