# HW1 - Principal Component Analysis

Phương Quỳnh Chi

December 27, 2022

**PCA algorithm - Maximum variance formulation**

Our goal is to project the data onto a space having dimensionality M < D while maximizing the variance of the projected data. We consider an independent and identically distributed (i.i.d.) dataset $X = x_1, x_2, ...x_n, x_n \in R^D$.

$$X = \begin{bmatrix} — & x_1^T & — \\ — & x_2^T & — \\ \vdots & \vdots & \vdots \\ — & x_n^T & — \end{bmatrix} \in R^{NxD}, \quad B = \begin{bmatrix} | & | & & | \\ b_1 & b_2 & ... & b_M \\ | & | & & | \end{bmatrix} \in R^{DxM}$$

$$\Rightarrow Z = X.B = \begin{bmatrix} x_1^T b_1 & x_1^T b_2 & ... & x_1^T b_n \\ x_2^T b_1 & x_2^T b_2 & ... & x_2^T b_n \\ \vdots & \vdots & ... & \vdots \\ x_n^T b_1 & x_n^T b_2 & ... & x_n^T b_n \end{bmatrix} \in R^{NxM}$$

We start by seeking a single vector $b_1 \in R^D$ that maximizes the variance of the projected data. So the variance of Z is given by:

$$\begin{aligned} V_Z &= \frac{1}{N}\sum_{n=1}^{N}(x_n^T b_1 - \mu_Z)^2 \\ &= \frac{1}{N}\sum_{n=1}^{N}(x_n^T b_1)^2 \quad \text{because } \mu_Z = 0 \\ &= \frac{1}{N}\sum_{n=1}^{N} b_1^T x_n x_n^T b_1 \\ &= b_1^T(\frac{1}{N}\sum_{n=1}^{N} x_n x_n^T) b_1 \\ &= b_1^T S b_1 \qquad \text{where S is covariance matrix.} \end{aligned}$$

Constrained optimization problem $\underset{b_1}{max} b_1^T S b_1$ subject to $||b_1||^2 = 1$.

To enforce this constraint, we introduce a Lagrange multiplier that we shall denote by $\lambda_1$, and then make an unconstrained maximization of:

$$L(b_1, \lambda_1) = b_1^T S b_1 + \lambda_1(1 - b_1^T b1)$$

$$\frac{\partial L}{\partial b_1} = 0 \Leftrightarrow 2Sb_1 - 2\lambda_1 b_1 = 0 \Leftrightarrow Sb_1 = \lambda b_1$$

$$\frac{\partial L}{\partial \lambda_1} = 0 \Leftrightarrow b_1^T b_1 = 1$$

We can see that $b_1, \lambda_1$ is an eigenvector and an eigenvalue of S respectively.

$$V = b_1^T S b_1 = b_1^T \lambda_1 b_1 = \lambda_1$$

The variance of the data projected onto a one-dimensional subspace equals the eigenvalue that is associated with the basis vector $b_1$ that spans this subspace.