ML2 - Hw1: PCA

Ngo Doan Kien - 11201976

December 2022

1 Problem 1

We consider an independent and identically distributed (i.i.d.) dataset $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ with mean 0, $\mathbf{x}_i \in \mathbb{R}^D$.

$$\mathbf{X} = \begin{bmatrix} -\mathbf{x}_1^T- \\ -\mathbf{x}_2^T- \\ \vdots \\ -\mathbf{x}_n^T- \end{bmatrix} \in \mathbb{R}^{NxD}$$

We assume there exists a low-dimensional compressed representation:

$$\mathbf{Z} = \mathbf{XB} \tag{1}$$

with projection matrix

$$\mathbf{B} = \begin{bmatrix} | & | & & | \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_M \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{DxM}$$

$$\begin{aligned}(1) \Leftrightarrow \begin{bmatrix} -\mathbf{z}_1^T- \\ -\mathbf{z}_2^T- \\ \vdots \\ -\mathbf{z}_n^T- \end{bmatrix} &= \begin{bmatrix} -\mathbf{x}_1^T- \\ -\mathbf{x}_2^T- \\ \vdots \\ -\mathbf{x}_n^T- \end{bmatrix} \cdot \begin{bmatrix} | & | & & | \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_M \\ | & | & & | \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{x}_1^T\mathbf{b}_1 & \mathbf{x}_1^T\mathbf{b}_2 & \cdots & \mathbf{x}_1^T\mathbf{b}_M \\ \mathbf{x}_2^T\mathbf{b}_1 & \mathbf{x}_2^T\mathbf{b}_2 & \cdots & \mathbf{x}_2^T\mathbf{b}_M \\ \vdots & \vdots & & \vdots \\ \mathbf{x}_N^T\mathbf{b}_1 & \mathbf{x}_N^T\mathbf{b}_2 & \cdots & \mathbf{x}_N^T\mathbf{b}_M \end{bmatrix}\end{aligned}$$

Consider $\mathbf{b}_1$ first, we have

$$\mu_X = \frac{\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_N}{N} = 0$$

$$\begin{aligned}\mu_Z &= \frac{\mathbf{x}_1^T\mathbf{b}_1 + \mathbf{x}_1^T\mathbf{b}_2 + \cdots + \mathbf{x}_1^T\mathbf{b}_M}{N} \\ &= \frac{1}{N}\mathbf{b}_1^T \sum_{i=N}^{N} \mathbf{x}_i \\ &= \mathbf{b}_1^T \mu_X = 0\end{aligned}$$

$$
\begin{aligned}
V_Z &= \frac{1}{N}\sum_{i=N}^{N}(\mathbf{x}_i^T\mathbf{b}_1 - \mu_Z)^2 \\
&= \frac{1}{N}\sum_{i=N}^{N}(\mathbf{x}_i^T\mathbf{b}_1)^2 \\
&= \frac{1}{N}\sum_{i=N}^{N}\mathbf{b}_1^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{b}_1 \\
&= \mathbf{b}_1^T(\frac{\sum_{i=N}^{N}\mathbf{x}_i\mathbf{x}_i^T}{N})\mathbf{b}_1 \\
&= \mathbf{b}_1^T S\mathbf{b}_1
\end{aligned}
$$

S is called the data covariance matrix.
Remark: increasing the magnitude of the vector $\mathbf{b}_1$ increases V. Therefore, we restrict all solutions to $||\mathbf{b}_1||^2 = 1$, which results in a constrained optimization problem in which we seek the direction along which the data varies most.
Constraint optimization problem:

$$\max_{\mathbf{b}_1} \mathbf{b}_1^T S\mathbf{b}_1$$

subject to

$$||\mathbf{b}_1||^2 = 1$$

The Lagrangian:

$$\mathcal{L}(\mathbf{b}_1, \lambda_1) = \mathbf{b}_1^T S\mathbf{b}_1 + \lambda_1(1 - \mathbf{b}_1^T\mathbf{b}_1)$$

$$\frac{\nabla\mathcal{L}}{\nabla\mathbf{b}_1} = 0 \Leftrightarrow S\mathbf{b}_1 = \lambda_1\mathbf{b}_1$$

$$\frac{\nabla\mathcal{L}}{\nabla\lambda_1} = 0 \Leftrightarrow \mathbf{b}_1^T S\mathbf{b}_1 = 1$$

$$V = \mathbf{b}_1^T S\mathbf{b}_1 = \mathbf{b}_1^T\lambda_1\mathbf{b}_1 = \lambda_1\mathbf{b}_1^T\mathbf{b}_1 = \lambda_1$$

The variance of the data projected onto a one-dimensional subspace equals the eigenvalue that is associated with the basis vector $\mathbf{b}_1$ that spans this subspace.

Therefore, to maximize the variance of the low-dimensional code, we choose the basis vector associated with the largest eigenvalue principal component of the data covariance matrix.

This eigenvector is called the first principal component.The second component is the projection of data onto the eigenvector corresponding with the second largest eigenvalue.