# PCA

ML Instruction Team, Fall 2022

CE Department
Sharif University of Technology

# Motivation

- PCA identifies one or more orthogonal directions that capture the greatest amount of variance in a feature matrix $X \in \mathbb{R}^{m \times n}$.

- Assuming zero-mean feature matrix $X \in \mathbb{R}^{m \times n}$, the variance of the samples' projections onto a unit vector $v$ is given by:

$$\frac{1}{n} \sum_{i=1}^{n} (x_i^t v)^2 = \frac{1}{n} \|Xv\|^2 = \frac{1}{n} v^t X^t X v$$

- In light of this consideration, we define the first desired vector $v_1$ as the solution to the constrained optimization problem:

$$\max_{\|v\|_2 = 1} v^t X^t X v$$

- We convert this constrained optimization problem into an unconstrained one by writing down its Lagrangian:

$$\mathcal{L}(v) := v^t X^t X v - \lambda(v^t v - 1)$$

# First PC

- First-order necessary conditions for optimal value imply that:

$$0 = \nabla \mathcal{L}(v_1) = 2X^t X v_1 - 2\lambda v_1$$

- Since $X^t X v_1 = \lambda v_1$, $v_1$ is an eigenvector of $X^t X$ with eigenvalue $\lambda$.

- Since we constrain $\|v_1\|_2^2 = v_1^t v_1 = 1$, the value of the objective is precisely:

$$v_1^t X^t X v_1 = v_1^t (\lambda v_1) = \lambda v_1^t v_1 = \lambda$$

- The optimal value is $\lambda = \lambda_{\max}(X^t X)$, which is achieved when $v_1$ is a unit eigenvector of $X^t X$ corresponding to its largest eigenvalue.

# More PCs?

- How to find more direction with the desired property?
  - ▷ Ideally, the subsequent directions found should also be directions of high variance.
  - ▷ They should be orthogonal to the existing ones in order to minimize redundancy.

- We define the $k$-th loading vector $v_k$ as the solution to the constrained optimization problem:

$$\max_{v} v^t X^t X v \quad \text{subject to} \quad v^t v = 1, v^t v_i = 0, \quad i = 1, \ldots, k - 1$$

- **Claim**: $v_k$ is a unit eigenvector of $X^t X$ corresponding to its $k$-th largest eigenvalue.

- The unit vector that defines the $k$-th axis is called the $k$-th principal component (PC).

# Evaluation of PCs

■ Assuming the singular value decompositon of feature matrix $X$ as follows:

$$X = U\Sigma V^T = [u_1, u_2, \cdots, u_r] \begin{bmatrix} \sigma_1 & & & 0 \\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_r \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_r^T \end{bmatrix}$$

The first $k$ PCs are $W_k = [v_1, v_2, \cdots, v_k]$.

■ Explained Variance Ratio explains the proportion of the dataset's variance that lies along the axis of each PC.

■ PCA can also be viewed as the projection of the sample points to the subspace with the minimum perpendicular distance.

# Other Derivation?

**Thank You!**

**Any Question?**