Principal Component Analysis

Dripta Mj
Department of Mathematics
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, India

Machine Learning
DA 220

Sem 2, 2019-20

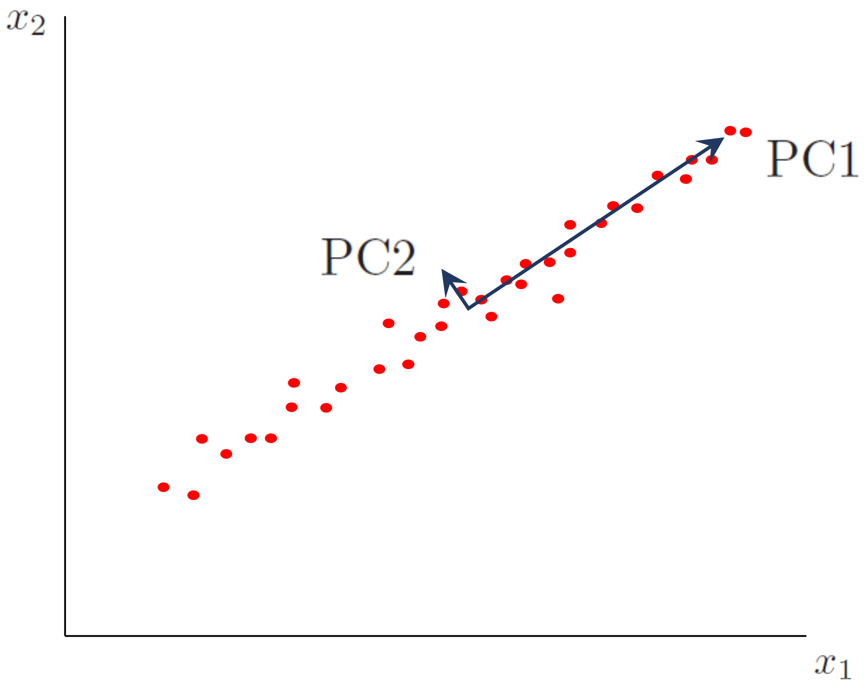
Dimensionality reduction

- Dimensionality reduction is the process of reducing the number of features of a dataset.
- Types: Feature selection, Feature extraction.
- Feature selection: Selects a subset of features.
  - Removes irrelevant features from the dataset.
- Feature extraction: Selects a few combinations of input features that capture most of the variations of the data.
  - Creates new features (through transformation) using existing ones.

Introduction to PCA

- Widely used method for dimensionality reduction.
- Original dataset – large number of interrelated input variables.
- Consider dataset: $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(N)}\}$, where $\mathbf{x}^{(n)}$is a D dimensional variable.
- Goal: Represent the data in a lower dimension Q $(< D)$.
  - Transform the data to a new uncorrelated set of variables – the principal components.
  - Extraction of the most informative Q linear combinations which explains the data.
  - This is the projection of the data in D dimensions onto a lower-dimensional subspace.
- Orthogonal projection of data onto a lower dimensional (linear) space, such that the variance of the projected data is maximized.

Principal components



- PC1: Direction of most variation
- PC2: Direction of second most variation orthogonal to PC1

Dataset

- Consider dataset: $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(N)}\}$, where $\mathbf{x}^{(n)}$is a D dimensional variable.

$$\mathbf{X} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & . & . & x_1^{(N)} \\ x_2^{(1)} & x_2^{(2)} & . & . & x_2^{(N)} \\ . & . & . & . & . \\ . & . & . & . & . \\ x_D^{(1)} & . & . & . & x_D^{(N)} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ . \\ . \\ \mathbf{x}_D \end{bmatrix}$$

- Want a lower-dimensional ($Q < D$) representation of the data:

$$\mathbf{Z} = \begin{bmatrix} z_1^{(1)} & z_1^{(2)} & . & . & z_1^{(N)} \\ . & . & . & . & . \\ z_Q^{(1)} & . & . & . & z_Q^{(N)} \end{bmatrix}$$

Variance

- Consider a vector $\mathbf{x} = [x_1, x_2, ..., x_N]$ having a mean value of 0.

- The variance of the vector $\mathbf{x}$ can be computed as

$$\sigma_{\mathbf{x}}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - 0)(x_i - 0)$$
$$= \frac{1}{N-1}\mathbf{x}\mathbf{x}^{\mathrm{T}}$$

Covariance

- Now consider two vectors: $\mathbf{x} = [x_1, x_2, ..., x_N]$ and $\mathbf{z} = [z_1, z_2, ..., z_N]$, both having mean 0.
- The covariance between vectors $\mathbf{x}$ and $\mathbf{z}$ can be computed as
$$\sigma^2_{\mathbf{xz}} = \frac{1}{N-1}\mathbf{x}\mathbf{z}^{\mathrm{T}}$$
  - Covariance measures the correlation between variables.
- If $\sigma^2_{\mathbf{xz}} \approx 0$ then $\mathbf{x}$ and $\mathbf{z}$ are almost uncorrelated.

Covariance matrix

- Assume data is centered.
- The covariance matrix $\mathbf{S}$ can be obtained as:

$$\mathbf{S} = \frac{1}{N-1}\mathbf{X}\mathbf{X}^{\mathrm{T}}.$$

- Can write the covariance matrix as

$$\mathbf{S} = \begin{bmatrix} \sigma^2_{\mathbf{x}_1} & \sigma^2_{\mathbf{x}_1\mathbf{x}_2} & \cdot & \cdot & \sigma^2_{\mathbf{x}_1\mathbf{x}_D} \\ \sigma^2_{\mathbf{x}_2\mathbf{x}_1} & \sigma^2_{\mathbf{x}_2} & \cdot & \cdot & \sigma^2_{\mathbf{x}_1\mathbf{x}_D} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \sigma^2_{\mathbf{x}_D\mathbf{x}_1} & \cdot & \cdot & \cdot & \sigma^2_{\mathbf{x}_D} \end{bmatrix}$$

- The i-th diagonal term corresponds to the variance in the i-th dimension of the problem.
- The off-diagonal terms are the covariances.
- Small off-diagonal term indicates that the variables are almost uncorrelated.
- $\mathbf{S}$ is symmetric.

Covariance matrix

- Want to transform the covariance matrix $\mathbf{S}$ to $\mathbf{S_Z}$ that has the following form:

$$\mathbf{S_Z} = \begin{bmatrix} \sigma^2_{\mathbf{Z}_1} & 0 & \cdot & \cdot & 0 \\ 0 & \sigma^2_{\mathbf{Z}_2} & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \sigma^2_{\mathbf{Z}_D} \end{bmatrix}$$

- The transformed matrix $\mathbf{S_Z}$ has no correlation between the different dimensions.
- Can order the variances such that: $\sigma^2_{\mathbf{Z}_1} \geq \sigma^2_{\mathbf{Z}_2} \geq .... \geq \sigma^2_{\mathbf{Z}_D}$.
- So $\sigma^2_{\mathbf{Z}_1}$ is the largest variance, and the dimension corresponding to it is known as the first principal component.
- Similarly $\sigma^2_{\mathbf{Z}_2}$ is the variance of the second principal component.

Eigenvalue decomposition

- Eigenvalue decomposition of the covariance matrix $\mathbf{S}$:

$$\mathbf{S} = \mathbf{V}\Lambda\mathbf{V}^{-1}$$

  where Λ is a diagonal matrix, $\mathbf{V}$ is a matrix of eigenvectors of $\mathbf{S}$ with columns corresponding to right eigenvectors of $\mathbf{S}$.

- The diagonal elements of Λ are the eigenvalues of $\mathbf{S}$ for the corresponding eigenvectors.

- Since $\mathbf{S}$ is symmetric, the eigenvalues are real and the eigenvectors are orthogonal to each other.

- The eigenvectors can be made orthonormal by taking $\mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{I}$.

Linear transformation

- Consider the following linear transformation of the original data $\mathbf{X}$ into $\mathbf{Z}$:

$$\mathbf{Z} = \mathbf{V}^{\mathrm{T}}\mathbf{X}$$

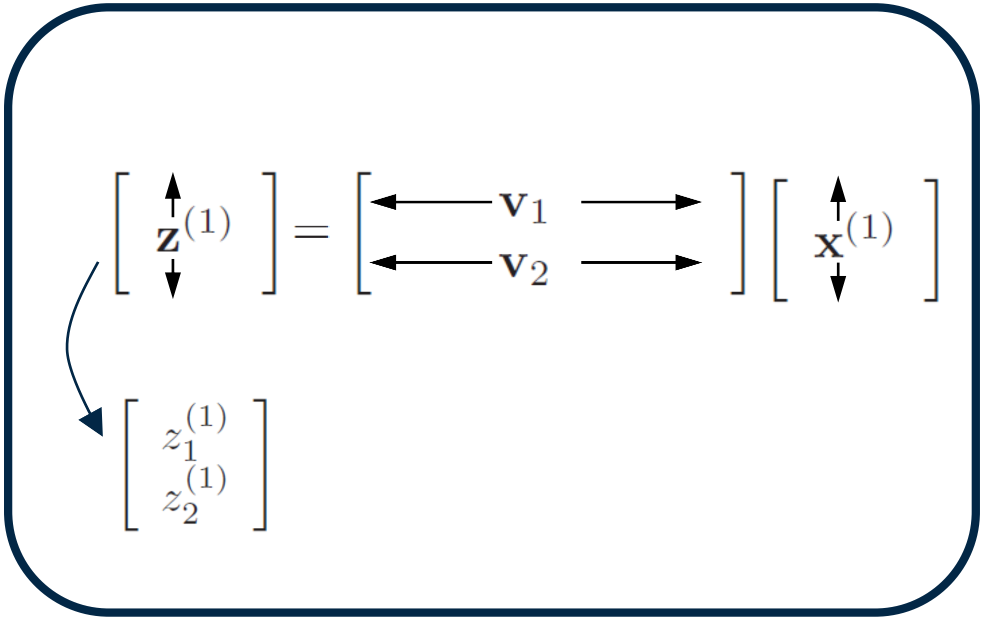
$$\begin{bmatrix} \uparrow & \uparrow & \cdots & \uparrow \\ & & \cdots & \\ \mathbf{z}^{(1)} & \mathbf{z}^{(2)} & \cdots & \mathbf{z}^{(N)} \\ & & \cdots & \\ \downarrow & \downarrow & \cdots & \downarrow \end{bmatrix} = \begin{bmatrix} \uparrow & \uparrow & \cdots & \uparrow \\ & & \cdots & \\ \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_D \\ & & \cdots & \\ \downarrow & \downarrow & \cdots & \downarrow \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \uparrow & \uparrow & \cdots & \uparrow \\ & & \cdots & \\ \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \cdots & \mathbf{x}^{(N)} \\ & & \cdots & \\ \downarrow & \downarrow & \cdots & \downarrow \end{bmatrix}$$

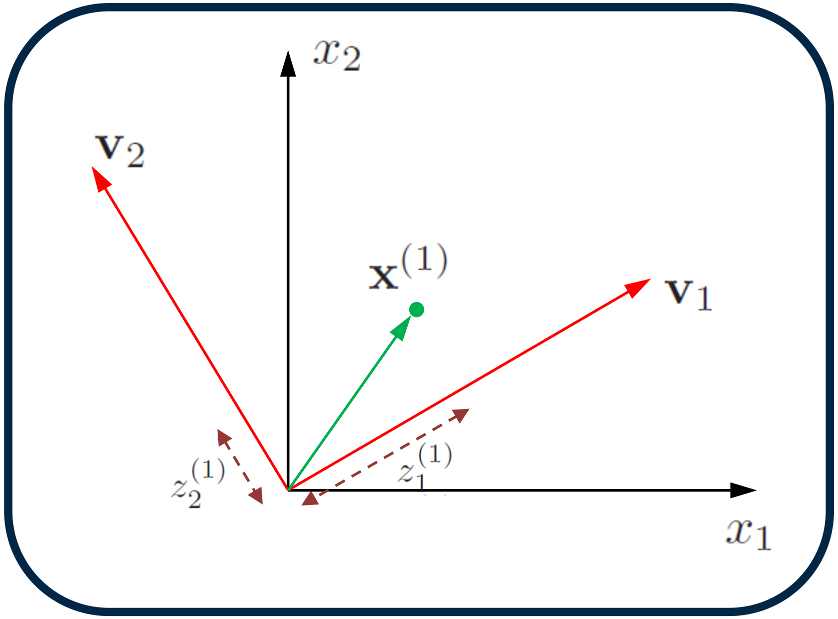
Linear transformation

- Consider the following linear transformation of the original data $\mathbf{X}$ into $\mathbf{Z}$:

$$\mathbf{Z} = \mathbf{V}^{\mathrm{T}}\mathbf{X}$$

- Consider a 2D example where the transformation is applied to a single data point $\mathbf{x}^{(1)}$





Linear transformation

- Consider the following linear transformation of the original data $\mathbf{X}$ into $\mathbf{Z}$:

$$\mathbf{Z} = \mathbf{V}^{\mathrm{T}}\mathbf{X}$$

- The covariance of $\mathbf{Z}$ can be obtained as:

$$\begin{aligned}
\mathbf{S}_{\mathbf{Z}} &= \frac{1}{N-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}} \\
&= \frac{1}{N-1}\left(\mathbf{V}^{\mathrm{T}}\mathbf{X}\right)\left(\mathbf{V}^{\mathrm{T}}\mathbf{X}\right)^{\mathrm{T}} \\
&= \frac{1}{N-1}\left(\mathbf{V}^{\mathrm{T}}\mathbf{X}\right)\left(\mathbf{X}^{\mathrm{T}}\mathbf{V}\right) \\
&= \frac{1}{N-1}\mathbf{V}^{\mathrm{T}}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\mathbf{V} \\
&= \mathbf{V}^{\mathrm{T}}\left(\frac{1}{N-1}\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\mathbf{V} \\
&= \mathbf{V}^{\mathrm{T}}\mathbf{S}\mathbf{V}
\end{aligned}$$

Covariance matrix

- Consider the following linear transformation of the original data $\mathbf{X}$ into $\mathbf{Z}$:

$$\mathbf{Z} = \mathbf{V}^{\mathrm{T}}\mathbf{X}$$

- The covariance of $\mathbf{Z}$ can be obtained as:

$$\begin{aligned}
\mathbf{S}_{\mathbf{Z}} &= \mathbf{V}^{\mathrm{T}}\mathbf{V}\Lambda\mathbf{V}^{-1}\mathbf{V} \\
&= (\mathbf{V}^{\mathrm{T}}\mathbf{V})\Lambda(\mathbf{V}^{\mathrm{T}}\mathbf{V}) \qquad (\mathbf{V}^{-1} = \mathbf{V}^{\mathrm{T}} \text{ as } \mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{I}\ ) \\
&= \Lambda
\end{aligned}$$

- The covariance matrix $\mathbf{S}_{\mathbf{Z}}$ is diagonal as Λ is diagonal.

Diagonal covariance matrix

- So we have

$$\mathbf{S_Z} = \Lambda = \begin{bmatrix} \sigma^2_{\mathbf{Z}_1} & 0 & . & . & 0 \\ 0 & \sigma^2_{\mathbf{Z}_2} & . & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & . & . & . & \sigma^2_{\mathbf{Z}_D} \end{bmatrix}$$

- The diagonal terms of $\mathbf{S_Z}$ correspond to variances along the dimensions of the transformed vector space.

- Note, the diagonal matrix Λ comprise the eigenvalues of $\mathbf{S}$.

- The variances along the projected dimensions (eigenvectors of $\mathbf{S}$) are the corresponding eigenvalues of $\mathbf{S}$.

METHOD

of

LAGRANGE MULTIPLIERS

Method of Lagrange multipliers

- The first principal component can be written as linear combination of the original variables as

$$\begin{aligned} z_1 &= v_{11}x_1 + v_{12}x_2 + ... + v_{1D}x_D \\ &= \mathbf{v}_1^{\mathrm{T}}\mathbf{x} \end{aligned}$$

where $\mathbf{v}_1^{\mathrm{T}} = \left[v_{11}, v_{21}, ..., v_{1D}\right]$.

- For the N given data points, the corresponding vector in the first dimension is given as

$$\mathbf{z}_1 = \mathbf{v}_1^{\mathrm{T}}\mathbf{X}.$$

- The variance in the first dimension is given as

$$\begin{aligned} \mathrm{var}(\mathbf{z}_1) &= \mathrm{var}(\mathbf{v}_1^{\mathrm{T}}\mathbf{X}) \\ &= \frac{1}{N-1}\mathbf{v}_1^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{v}_1 \\ &= \mathbf{v}_1^{\mathrm{T}}\mathbf{S}\mathbf{v}_1 \end{aligned}$$

and we want $\mathrm{var}(\mathbf{z}_1)$ to be maximized.

1^{st} principal component

- Maximize the projected variance $\mathbf{v}_1^T \mathbf{S} \mathbf{v}_1$ with respect to $\mathbf{v}_1$ subject to normalization constraint: $\mathbf{v}_1^T \mathbf{v}_1 = 1$.

- Approach: Use the method of Lagrange multiplier to find the maximum of an objective function subject to a constraint.

- Consider the Lagrangian: $\mathcal{L}_1 = \mathbf{v}_1^T \mathbf{S} \mathbf{v}_1 + \lambda_1 (1 - \mathbf{v}_1^T \mathbf{v}_1)$

  - Objective: $\max \mathcal{L}_1$
    - Differentiating $\mathcal{L}_1$ with respect to $\mathbf{v}_1$ and equating to 0:

$$\frac{\mathrm{d}\mathcal{L}_1}{\mathrm{d}\mathbf{v}_1} = \mathbf{S}\mathbf{v}_1 - \lambda_1 \mathbf{v}_1 = 0$$

1st principal component

- Therefore we have

$$\mathbf{S}\mathbf{v}_1 = \lambda_1 \mathbf{v}_1$$

  - λ_1 is an eigenvalue of $\mathbf{S}$, and $\mathbf{v}_1$ is the associated eigenvector.

- Multiplying both sides by $\mathbf{v}_1^T$, we have:

$$\mathbf{v}_1^T \mathbf{S}\mathbf{v}_1 = \lambda_1 \mathbf{v}_1^T \mathbf{v}_1$$
$$= \lambda_1$$

- Note that we want to maximize $\mathbf{v}_1^T \mathbf{S}\mathbf{v}_1$.

- Therefore λ_1 is the largest eigenvalue of $\mathbf{S}$. This is called the 1st principal component.

2nd principal component

- The second principal component too can be written as linear combination of the original variables as

$$\begin{aligned} z_2 &= v_{21}x_1 + v_{22}x_2 + ... + v_{2D}x_D \\ &= \mathbf{v}_2^{\mathrm{T}}\mathbf{x} \end{aligned}$$

where $\mathbf{v}_2^{\mathrm{T}} = \left[v_{21}, v_{22}, ..., v_{2D}\right]$.

- The projection of the N data points in the second dimension can be given as

$$\mathbf{z}_2 = \mathbf{v}_2^{\mathrm{T}}\mathbf{X}.$$

- Want $\mathbf{z}_2$ to be uncorrelated to $\mathbf{z}_1$ i.e.

$$\text{covariance}\left(\mathbf{z}_1, \mathbf{z}_2\right) = 0$$

therefore we have

$$\frac{1}{N-1}\mathbf{v}_1^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{v}_2 = 0 \;\Rightarrow\; \mathbf{v}_1^{\mathrm{T}}\left(\frac{1}{N-1}\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\mathbf{v}_2 = 0$$

$$\Rightarrow \mathbf{v}_1^{\mathrm{T}}\mathbf{S}\mathbf{v}_2 = 0 \Rightarrow \mathbf{v}_2^{\mathrm{T}}\mathbf{S}\mathbf{v}_1 = 0 \Rightarrow \mathbf{v}_2^{\mathrm{T}}\lambda_1\mathbf{v}_1 = 0 \Rightarrow \lambda_1\mathbf{v}_2^{\mathrm{T}}\mathbf{v}_1 = 0 \Rightarrow \boxed{\mathbf{v}_2^{\mathrm{T}}\mathbf{v}_1 = 0}$$

- Objective: $\boxed{\max \mathbf{v}_2^{\mathrm{T}}\mathbf{S}\mathbf{v}_2 \quad \text{such that} \quad \mathbf{v}_2^{\mathrm{T}}\mathbf{v}_2 = 1 \quad \text{and} \quad \mathbf{v}_2^{\mathrm{T}}\mathbf{v}_1 = 0}$

2^{nd} principal component

- Construct the Lagrangian:

$$\mathcal{L}_2 = \mathbf{v}_2^T \mathbf{S} \mathbf{v}_2 + \lambda_2 (1 - \mathbf{v}_2^T \mathbf{v}_2) + \mu_1 (\mathbf{v}_2^T \mathbf{v}_1)$$

- Objective: $\max \ \mathcal{L}_2$
  - Differentiating $\mathcal{L}_2$ with respect to $\mathbf{v}_2$ and equating to 0:

$$\frac{\mathrm{d}\mathcal{L}_2}{\mathrm{d}\mathbf{v}_2} = 2\mathbf{S}\mathbf{v}_2 - 2\lambda_2 \mathbf{v}_2 + \mu_1 \mathbf{v}_1 = 0$$

  - Multiplying both sides by $\mathbf{v}_1^T$:

$$2\mathbf{v}_1^T \mathbf{S} \mathbf{v}_2 - 2\lambda_2 \mathbf{v}_1^T \mathbf{v}_2 + \mu_1 \mathbf{v}_1^T \mathbf{v}_1 = 0$$

  - Using $\mathbf{v}_1^T \mathbf{v}_1 = 1$, and $\mathbf{v}_1^T \mathbf{v}_2 = 0$ for $\max \ \mathcal{L}_2$, we have

$$2\mathbf{v}_1^T \mathbf{S} \mathbf{v}_2 + \mu_1 = 0$$

2nd principal component

- Now we have

$$\mathbf{v}_1^T \mathbf{S} \mathbf{v}_2 = \mathbf{v}_2^T (\mathbf{S} \mathbf{v}_1) = \mathbf{v}_2^T (\lambda_1 \mathbf{v}_1) = \lambda_1 (\mathbf{v}_2^T \mathbf{v}_1) = 0,$$

and on substitution in ■ gives

$$\mu_1 = 0.$$

- On substitution of $\mu_1 = 0$ in ■ yields

$$\mathbf{S} \mathbf{v}_2 = \lambda_2 \mathbf{v}_2.$$

- Therefore λ_2 is another eigenvalue of $\mathbf{S}$.

- Since we want to maximize $\mathbf{v}_2^{\mathrm{T}} \mathbf{S} \mathbf{v}_2$ $(= \mathbf{v}_2^{\mathrm{T}} \lambda_2 \mathbf{v}_2 = \lambda_2)$ and also want $\mathbf{v}_2$ to be uncorrelated to $\mathbf{v}_1$, λ_2 should be the second largest eigenvalue of $\mathbf{S}$.

Percentage of variance

- The percentage of variance explained by the jth principal component:

$$PV_j = \frac{\lambda_j}{\sum_{i=1}^{D} \lambda_i} \times 100$$

- The percentage of variance accounted for by the first Q principal components is given by:

$$PV = \frac{\sum_{i=1}^{Q} \lambda_i}{\sum_{i=1}^{D} \lambda_i} \times 100$$