Principal Component Analysis (PCA)

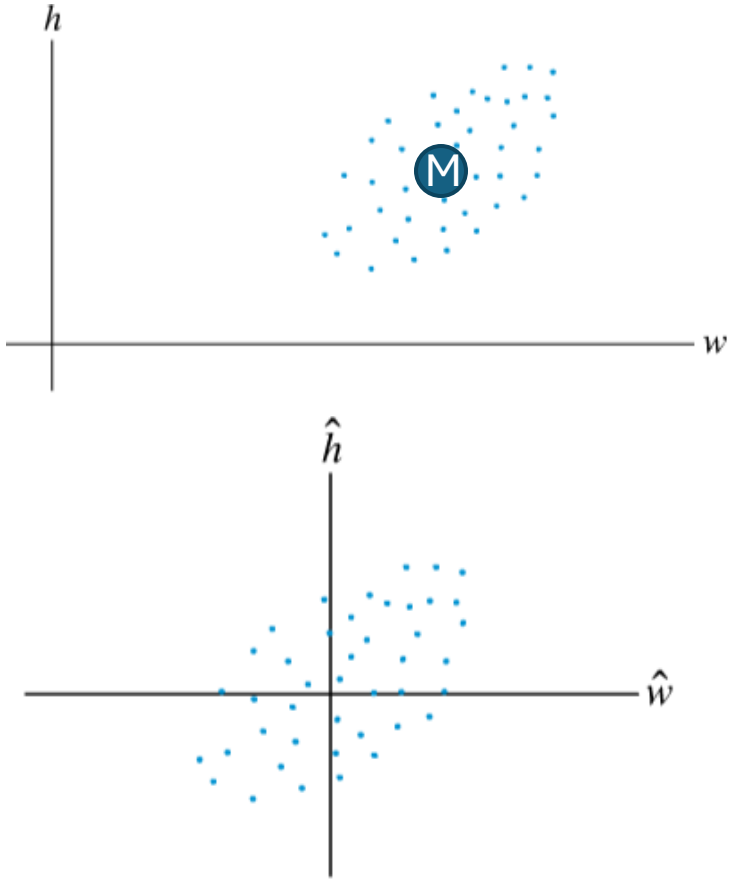
How to Get the Principal Components?

- The first step is to **center** the data points.
- i.e. subtract the mean of each attribute from the corresponding coordinate.
- Example: Consider the matrix of observations:

$$\begin{bmatrix} w_1 & w_2 & \cdots & w_N \\ h_1 & h_2 & \cdots & h_N \end{bmatrix}$$

$$\uparrow \quad \uparrow \qquad \uparrow$$

$$\mathbf{X}_1 \quad \mathbf{X}_2 \qquad \mathbf{X}_N$$



- The sample mean, M is given by

$$\mathbf{M} = \frac{1}{N}(\mathbf{X}_1 + \cdots + \mathbf{X}_N)$$

- The sample mean is the point in the center.
- For k = 1…N, let $\hat{\mathbf{X}}_k = \mathbf{X}_k - \mathbf{M}$
- The columns of B have a zero sample mean $B = [\hat{\mathbf{X}}_1 \quad \hat{\mathbf{X}}_2 \quad \cdots \quad \hat{\mathbf{X}}_N]$
- B is said to be in **mean-deviation** form.

Example

- Three measurements are made on each of four individuals in a random sample from a population.
- The observation vectors are:

$$\mathbf{X}_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \quad \mathbf{X}_2 = \begin{bmatrix} 4 \\ 2 \\ 13 \end{bmatrix}, \quad \mathbf{X}_3 = \begin{bmatrix} 7 \\ 8 \\ 1 \end{bmatrix}, \quad \mathbf{X}_4 = \begin{bmatrix} 8 \\ 4 \\ 5 \end{bmatrix}$$

- Determine the coordinate of the centered data.

```
X = np.array([[1,4,7,8],[2,2,8,4],[1,13,1,5]])
X
```

```
array([[ 1,  4,  7,  8],
       [ 2,  2,  8,  4],
       [ 1, 13,  1,  5]])
```

```
# Find mean of each row
m=X.mean(axis = 1)
m
```

```
array([5., 4., 5.])
```

```
# Convert to a column array
m = m.reshape(-1,1)
m
```

```
array([[5.],
       [4.],
       [5.]])
```

```
B = X-m
B
```

```
array([[-4., -1.,  2.,  3.],
       [-2., -2.,  4.,  0.],
       [-4.,  8., -4.,  0.]])
```

How to Get the Principal Components? – cont.

- The second step is to find the **covariance matrix** for the *d* features.
- A $d \times d$ covariance matrix will look like $\begin{bmatrix} var(x_1) & \cdots & cov(x_1, x_d) \\ \vdots & \ddots & \vdots \\ cov(x_d, x_1) & \cdots & var(x_d) \end{bmatrix}$
- The main diagonal of the covariance matrix, contains the variances. e.g. $var(x_1)$ denotes how spread out the data are along 1st dimension.
- The off diagonal elements indicates if features change together (i.e. if x_1 increases x_2 increases) or in opposite direction (i.e. if x_1 increases x_2 decreases)
- The covariance matrix is symmetric.

Example

- The sample covariance matrix of a data set is as follows: $\begin{bmatrix} 10 & 6 & 0 \\ 6 & 8 & -8 \\ 0 & -8 & 32 \end{bmatrix}$
- Interpret the numbers.
  1. Since the covariance matrix is 3 by 3, the data has a dimension of 3.
  2. The entries in the third dimension has the widest spread of values compare to the first and second dimensions.
  3. The first and second dimensions are positively correlated.
  4. The second and third dimensions are negatively correlated.
  5. The first and third dimensions are **uncorrelated.**
- Analysis of the multivariate data is greatly simplified when most or all of the variables are uncorrelated, that is, when the **covariance matrix of the data is diagonal** or nearly diagonal.

Calculating the Sample Covariance Matrix

- $sample\ cov(x_1, x_2) = \frac{1}{N-1}\sum_{i=1}^{N}(x_{1i} - m_1)(x_{2i} - m_2)$
- Since we already centered the data, m_1 and m_2 are zero
- $sample\ cov(x_1, x_2) = \frac{1}{N-1}\sum_{i=1}^{N} x_{1i}x_{2i}$
- What's the covariance expression in matrix form?
- Let B be the centered data points: $B = \begin{bmatrix} x_{11} - m_1 & \cdots & x_{1N} - m_1 \\ \vdots & \ddots & \vdots \\ x_{d1} - m_d & \cdots & x_{dN} - m_d \end{bmatrix}$
- Then $BB^T = \begin{bmatrix} x_{11} - m_1 & \cdots & x_{1N} - m_1 \\ \vdots & \ddots & \vdots \\ x_{d1} - m_d & \cdots & x_{dN} - m_d \end{bmatrix} \begin{bmatrix} x_{11} - m_1 & \cdots & x_{d1} - m_d \\ \vdots & \ddots & \vdots \\ x_{1N} - m_1 & \cdots & x_{dN} - m_d \end{bmatrix}$
- Therefore sample covariance will be:

$$s = \frac{1}{N-1}BB^T$$

Example

- Three measurements are made on each of four individuals in a random sample from a population.
- The observation vectors are:

$$\mathbf{X}_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \quad \mathbf{X}_2 = \begin{bmatrix} 4 \\ 2 \\ 13 \end{bmatrix}, \quad \mathbf{X}_3 = \begin{bmatrix} 7 \\ 8 \\ 1 \end{bmatrix}, \quad \mathbf{X}_4 = \begin{bmatrix} 8 \\ 4 \\ 5 \end{bmatrix}$$

- Determine the covariance matrix.

```
N = B.shape[1]
N
```

```
4
```

```
sigma = B@B.T/(N-1)
sigma
```

```
array([[10.,  6.,  0.],
       [ 6.,  8., -8.],
       [ 0., -8., 32.]])
```

How to Get the Principal Components? – cont.

- The **eigenvector of the covariance matrix** with the largest eigenvalue is the direction along which the data set has the **maximum variance**.
- Therefore, final step in finding the principal components is to find the eigenvectors of the covariance matrix.

```
from numpy import linalg
l, ev = linalg.eig(sigma)
l
```

```
array([ 1.60571114, 13.84296424, 34.55132462])
```

```
ev
```

```
array([[-0.56861003, -0.8192675 , -0.07404999],
       [ 0.79551281, -0.52473595, -0.30300421],
       [ 0.20938481, -0.23119895,  0.95010791]])
```