Principal Component Analysis

1. Statistical Properties

Before talking about Principal Component Analysis (PCA), let's first review some statistical properties in the form of three different examples.

1.1. **Example 1: Height vs. Weight.**

In this example, we draw the heights and weights from a randomly selected population. In this case, the height and weights of each person are, in general, positively correlate. Selecting a random set of n samples, we can define $\mathbf{X}$ as an $m \times n$ matrix where $m = 2$, the number properties measured (in this case, height and weight). $\mathbf{X}$ is then defined as:

$$\mathbf{X} = \begin{bmatrix} w_1 & w_2 & \cdots & w_n \\ h_1 & h_2 & \cdots & h_n \end{bmatrix}$$

The statistical properties of this data set are as follows:

Averages.

$$\bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i \quad \text{(Average Weight)}$$

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_i \quad \text{(Average Height)}$$

Variance.

$$S_w^2 = \frac{1}{n-1}\sum_{i=1}^{n}(w_i - \bar{w})^2$$

$$S_h^2 = \frac{1}{n-1}\sum_{i=1}^{n}(h_i - \bar{h})^2$$

Covariance.

$$S_{wh}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(w_i - \bar{w})(h_i - \bar{h})$$

where S_{wh} positive means w large is correlated with h large.

1.2. **Example 2: Rainy Days in Summer vs. Ice Cream Sales.**

In this example, we look at the total sales of ice cream as well as the number of rainy days in a given summer. In this case, the sales of ice cream and number of rainy days are generally negatively correlated. We then define $\mathbf{X}$ again as an $m \times n$ matrix where $m = 2$ and n randomly selected samples. $\mathbf{X}$ can then be defined as:

$$\mathbf{X} = \begin{bmatrix} r_1 & r_2 & \cdots & r_n \\ s_1 & s_2 & \cdots & s_n \end{bmatrix}$$

Covariance can again be defined as

$$S_{rs} = \frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})$$

and S_{rs} is negative, representing the negative correlation between large r and large s.

Before we look at the 3rd example, let's briefly review the Multivariate Normal Distribution.

1.3. **Review of Multivariate Normal Distribution.** Let $\mathbf{x} \in \mathbb{R}^2$ be a random variable. We say that x has a multivariate normal distribution if for every vector $\mathbf{u}$, the scalar random variable $\mathbf{ux}$ has a normal (Gaussian) distribution. When this holds, the probability density function takes the form

$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}} \frac{1}{|\det \Sigma|} \exp(-\frac{1}{2}(\mathbf{x}-\mathbf{u})^* \mathbf{\Sigma}^* (\mathbf{x}-\mathbf{u}))$$

where $\mathbf{u} \in \mathbb{R}^2$ is the mean and $\mathbf{\Sigma}$ is a 2×2 pd matrix known as the covariance matrix.

Let $\mathbf{Z} = \left[\begin{array}{cccc} \mathbf{z}^{(1)} & \mathbf{z}^{(2)} & \cdots & \mathbf{z}^{(n)} \end{array} \right]$ be an $m \times n$ matrix of samples. Then the empirical mean is

$$\mathbf{z}_i = \frac{1}{n} \sum_{j=1}^{n} z_{ij} \quad \text{and} \quad \mathbf{z} = \begin{bmatrix} \bar{z}_1 \\ \vdots \\ \bar{z}_n \end{bmatrix}$$

Then $\mathbf{z}$ is an estimate for $\boldsymbol{\mu}$, the equilibrium point. Set

$$\mathbf{X} = \mathbf{Z} - \mathbf{z} \underbrace{\left[\begin{array}{cccc} 1 & 1 & \cdots & 1 \end{array} \right]}_{1 \times n}$$

$$\mathbf{S} = \frac{1}{n-1} \mathbf{XX}^*$$

We call $\mathbf{S}$ the empirical covariance matrix, which is an estimate for $\mathbf{\Sigma}$.

In PCA, make an assumption that the underlying data comes from a multivariate normal distribution.

1.4. **Example 3: Moving Spring Mass System.** In this example, we look at a mass spring system moving back and forth in a 3D system. At times $t_1, t_2, \ldots, t_n$, we record the position $\mathbf{z}^{(j)}$ of the mass to obtain a matrix $\mathbf{Z}$ such that $\mathbf{Z}$ is $m \times n$, $m = 3$, and defined as:

$$\mathbf{Z} = \left[\begin{array}{cccc} \mathbf{z}^{(1)} & \mathbf{z}^{(2)} & \cdots & \mathbf{z}^{(n)} \end{array} \right]$$

In other words,

$$\mathbf{z}^{(j)} = \boldsymbol{\mu} + \cos(\omega t_j) A\mathbf{u} + \mathbf{n}^{(j)}$$

where $\boldsymbol{\mu}$ is the equilibrium point, $\cos(\omega t_j) A\mathbf{u}$ represents the motion of the mass and $\mathbf{n}^{(j)}$ represents the noise. Set

$$\mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix} \quad \text{and} \quad m_j = \frac{1}{n} \sum_{j=1}^{n} z_{ij}$$

Then $\mathbf{m} \approx \boldsymbol{\mu}$. There are now 3 covariances and 3 variances. Subtract the average value from each row

$$\mathbf{X} = \mathbf{Z} - \mathbf{m} \underbrace{\left[\begin{array}{cccc} 1 & 1 & \cdots & 1 \end{array} \right]}_{1 \times n} = \begin{bmatrix} z_{11} - m_1 & z_{12} - m_1 & \cdots \\ z_{21} - m_2 & z_{22} - m_2 & \cdots \\ z_{31} - m_3 & \cdots & \ddots \end{bmatrix}$$

We then define

$$\mathbf{S} = \frac{1}{n-1} \mathbf{XX}^* = \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{bmatrix}$$

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} x_{ik} x_{kj} = \frac{1}{n-1} \sum_{k=1}^{n} (z_{ik} - m_i)(z_{jk} - m_j)$$

where $\mathbf{S}$ is the Empirical Covariance Matrix of size 3×3.

REFERENCES