Lecture - Pricipal Component Analysis

Dr. Arabin Kumar Dey

Assistant Professor
Department of Mathematics
Indian Institute of Technology Guwahati

April 12, 2012

Outline

Outline

- Consider weekly rates of return for five stocks (Allied Chemical, du Pont, Union Carbide, Exxon and Texaco) listed on the New York Stock Exchange were determined for the period January 1975 tnhrough December 1976. The weekly rates of return are defined as (current Friday closing price - previous Friday closing price)/(previous Friday closing price) adjusted for stock splits and dividends.

- In the above context we may be interested to know about most dominant stocks creating the overall market variation.
- Can we infer about most effective weeks in terms of the outcome of all stock returns ?
- Can we say something about the correlation between general stock market behavior or market component and industry activity or some other factors ?

Outline

Let the random vector have $X' = \{X_1, X_2, \cdots, X_p\}$ have covariance matrix Σ with eigen values $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p$. Consider the linear combinations :

$$\begin{aligned} Y_1 &= l_1'X = l_{11}X_1 + l_{21}X_2 + \cdots + l_{p1}X_p \\ Y_2 &= l_2'X = l_{12}X_1 + l_{22}X_2 + \cdots + l_{p2}X_p \\ \vdots &= \quad \vdots \\ Y_p &= l_p'X = l_{1p}X_1 + l_{2p}X_2 + \cdots + l_{pp}X_p \end{aligned}$$

$Var(Y_i) = l_i'\Sigma l_i, \quad i = 1, 2, \cdots, p$
$Cov(Y_i, Y_k) = l_i'\Sigma l_k, \quad i, k = 1, 2, \cdots, p$

First Principal Component = linear combination $l_1^{'}X$ that maximizes $Var(l_1^{'}X)$ subject to $l_1^{'}l_1 = 1$.

Second Principal component= linear combination $l_2^{'}X$ that maximizes $Var(l_2^{'}X)$ subject to $l_2^{'}l_2 = 1$ and $Cov(l_1^{'}X, l_2^{'}X) = 0$

i-th Principal component = linear combination $l_i^{'}X$ that maximizes $Var(l_i^{'}X)$ subject to $l_i^{'}l_i = 1$ and $Cov(l_i^{'}X, l_k^{'}X) = 0$ for $k < i$

Let Σ be the covariance matrix associated with the random vector $X^{'} = [X_1, \cdots, X_p]$. Let Σ have the eigenvalue-eigenvector pairs (λ_i, e_i) where $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. The ith principal component is given by

$$Y_i = e_i^{'} X = e_{1i} X_1 + e_{2i} X_2 + \cdots + e_{pi} X_p$$

With these choice,
$Var(Y_i) = e_i^{'} \Sigma e_i = \lambda_i, \quad i = 1, 2, \cdots, p$
$Cov(Y_i, Y_k) = e_i^{'} \Sigma e_k = 0, \quad i \neq k$

Maximization of Quadratic Forms for points on the Unit Sphere:

Let $B_{p\times p}$ be a positive definite matrix with eigen values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ and associated normalized eigenvectors $e_1, e_2, \cdots, e_p$. Then $\max_{x\neq 0} \frac{x^{'}Bx}{x^{'}x} = \lambda_1$ attained when $x = e_1$ $\max_{x\neq 0} \frac{x^{'}Bx}{x^{'}x} = \lambda_p$ attained when $x = e_p$ $\max_{x\perp e_1,\cdots,e_k} \frac{x^{'}Bx}{x^{'}x} = \lambda_{k+1}$ attained when $x = e_{k+1}, k = 1, 2, \cdots, (p-1)$ where symbol $\perp$ is read "perpendicular to".

Proof :

Let $B^{1/2} = P\Lambda^{1/2}P^{'}$ and $y = P^{'}x$. Consequently, $x \neq 0$ implies $y \neq 0$. Thus

$$\frac{x^{'}Bx}{x^{'}x} = \frac{x^{'}B^{1/2}B^{1/2}x}{x^{'}PP^{'}x} = \frac{y^{'}\Lambda y}{y^{'}y} = \frac{\sum_{i=1}^{p}\lambda_i y_i^2}{\sum_{i=1}^{p} y_i^2} \leq \lambda_1 \frac{\sum_{i=1}^{p} y_i^2}{\sum_{i=1}^{p} y_i^2} = \lambda_1$$

Setting $x = e_1$, $y = P^{'}e_1 = e_1$ and

$$\begin{cases} e_k^{'}e_1 = 1, & \text{if } k = 1 \\ 0, & \text{otherwise} \end{cases}$$

Proof : continues

For this choice of x, $\frac{y^{'}\Lambda y}{y^{'}y} = \lambda_1$.
Now, $x = Py = y_1e_1 + y_2e_2 + \cdots + y_pe_p$, so $x \perp e_1, e_2, \cdots, e_k$ implies

$$0 = e_i^{'}x = y_1e_i^{'}e_1 + y_2e_i^{'}e_2 + \cdots + y_pe_i^{'}e_p = y_i, i \leq k$$

Therefore $\frac{x^{'}Bx}{x^{'}x} = \frac{\sum_{i=k+1}^{p} \lambda_i y_i^2}{\sum_{i=k+1}^{p} y_i^2} \leq \lambda_{k+1}$

Objective

1. It is used for reducing the dimension of the data.
2. It can be used to rank the individuals.
3. It can be used as a tool for clustering.

Outline

Suppose the random variables X_1, X_2 and X_3 have the variance covariance matrix

$$\begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 11 \end{pmatrix}$$

It may be varified that the eigenvalue-eigenvector pairs are

$\lambda_1 = 5.83,\ e_1^{'} = [0.383, -0.924, 0]$

$\lambda_2 = 2.00,\ e_2^{'} = [0, 0, 1]$

$\lambda_3 = 0.17,\ e_3^{'} = [0.924, 0.383, 0]$

Example

Therefore, the principal components become

$$Y_1 = e_1^{'}X = 0.383X_1 - 0.924X_2$$

$$Y_2 = e_2^{'}X = X_3$$

$$Y_3 = e_3^{'}X = 0.924X_1 + 0.383X_2$$

The variable X_3 is one of the principal components because it is uncorrelated with the other two variables.

$$\begin{aligned} & Var(Y_1) = Var(0.383X_1 - 0.924X_2) \\ = & \ (0.383)^2 Var(X_1) + (-0.924)^2 Var(X_2) + 2(0.383)(-0.924)Cov(X_1, \\ = & \ 0.147(1) + 0.854(5) - 0.708(-2) \\ = & \ 5.83 = \lambda_1 \end{aligned}$$

$$
\begin{aligned}
Cov(Y_1, Y_2) &= Cov(0.383X_1 - 0.924X_2, X_3) \\
&= 0.383Cov(X_1, X_3) - 0.924Cov(X_2, X_3) \\
&= 0.383(0) - 0.924(0) = 0
\end{aligned}
$$

It is also readily apparent that

$$\sigma_{11} + \sigma_{22} + \sigma_{33} = 1 + 5 + 2 = \lambda_1 + \lambda_2 + \lambda_3 = 5.83 + 2.00 + 0.17$$

- The proportion of total variance accounted for by the first principal component is $\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} = 5.83/8 = 0.73$.
- Continuing the first two components account for a proportion $(5.83 + 2)/8 = 0.98$ of the population variance. In this case the components Y_1 and Y_2 could replace the three original variables with little loss of information.