MULTIVARIATE STATISTICAL ANALYSIS

Lecture 4
Principal Component Analysis

Associate Professor Lý Quốc Ngọc

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

fit@hcmus

Contents

4.1. Purpose of PCA

Principal component analysis is a statistical method to discover the **variance-covariance structure** of a set of variables through a **linear combination** of these variables in order to **reduce the data dimension** and **data interpretation**.

4.2. Problem statement

fit@hcmus

Let the random vector $X' = [X_1, X_2, ..., X_p]$ have the covariance matrix Σ .

Consider the linear combinations:

$$Y_1 = a_1' X = a_{11} X_1 + a_{12} X_2 + \ldots + a_{1p} X_p$$

$$Y_2 = a_2' X = a_{21} X_1 + a_{22} X_2 + \ldots + a_{2p} X_p$$

$$\vdots$$

$$Y_p = a_p' X = a_{p1} X_1 + a_{p2} X_2 + \ldots + a_{pp} X_p$$

We obtain: $$Var(Y_i) = a_i' \Sigma a_i, \ \mathrm{i} = 1,2,...,\mathrm{p}$$

$$Cov(Y_i, Y_j) = a_i' \Sigma a_j, \mathrm{i}, \mathrm{j} = 1,2,...,\mathrm{p}$$

Estimate a_i to have the principal components Y_i are those uncorrelated linear combinations whose variances are as large as possible.

4.3. Method

Estimate a_i to have the principal components Y_i are those uncorrelated linear combinations whose variances are as large as possible :

$$a_i^* = \underset{a_i}{\arg\max}(Var(Y_i)) = \underset{a_i}{\arg\max}(a_i^{'}\Sigma a_i),\ i = 1,2,\ldots,p$$

$$\|a_i\|_2^2 = a_i^{'} a_i = 1$$

$$Cov(Y_i, Y_j) = 0 \ \ \forall \mathrm{i} \neq \mathrm{j}$$

4.3. Method

Consider Lagrange function:

$$L(a_i,\lambda) = a_i^{'}\sum a_i + \lambda(1 - a_i^{'}a_i),\ i = 1,2,\ldots,p$$

The necessary condition for $Var(Y_i)$ reaching the maximum with constraints is :

$$\frac{\partial L(a_i,\lambda)}{\partial a_i} = 0$$

$$\Rightarrow 2\sum a_i - 2\lambda a_i = 0,\ i = 1,2,\ldots,p$$

$$\Rightarrow \sum a_i = \lambda a_i$$

So a_i is the eigenvector corresponding to the eigenvalues λ of $\sum$ and

$$Cov(Y_i,Y_j) = a_i^{'}\sum a_j = a_i^{'}\lambda a_j = 0,\ \forall \mathrm{i} \neq \mathrm{j}$$

4.4. Geometrical Explanation

Discuss in class

4.5. Model Checking

Discuss in class

4.6. Case study

Discuss in class