Lecture 9

Principle Components Analysis

Reading: Zelterman Chapter 8.1-8.4; DSA 8020 Lecture 12 [Link]

DSA 8070 Multivariate Analysis
October 11-October 15, 2021

Whitney Huang
Clemson University

Agenda

History

- First introduced by Karl Pearson (1901) as a procedure for finding lines and planes which best fit a set of points in p-dimensional space

LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* By KARL PEARSON, *F.R.S., University College, London* *.

(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane. Analytically this consists in taking

$$y = a_0 + a_1 x, \quad \text{or} \quad z = a_0 + a_1 x + b_1 y,$$
$$\text{or} \quad z = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots + a_n x_n,$$

where $y, x, z, x_1, x_2, \ldots x_n$ are variables, and determining the "best" values for the constants $a_0, a_1, b_1, a_0, a_1, a_2, a_3, \ldots a_n$ in relation to the observed corresponding values of the variables. In nearly all the cases dealt with in the text-books of least squares, the variables on the right of our equations are treated as the independent, those on the left as the dependent variables. The result of this treatment is that we get one straight line or plane if we treat some one variable as independent, and a quite different one if we treat another variable as the independent variable. There is no paradox

- Harold Hotelling (1933) published a paper on PCA to find a smaller "fundamental set of independent variables" that determines the values of the original set of p variables

ANALYSIS OF A COMPLEX OF STATISTICAL VARIABLES INTO PRINCIPAL COMPONENTS[1]

HAROLD HOTELLING

Columbia University

1. INTRODUCTION

Consider n variables attaching to each individual of a population. These statistical variables $x_1, x_2, \ldots, x_n$ might for example be scores made by school children in tests of speed and skill in solving arithmetical problems or in reading; or they might be various physical properties of telephone poles, or the rates of exchange among various currencies. The x's will ordinarily be correlated. It is natural to ask whether some more fundamental set of independent variables exists, perhaps fewer in number than the x's, which determine the values the x's will take. If $\gamma_1, \gamma_2, \ldots$ are such variables, we shall then have a set of relations of the form

$$x_i = f_i(\gamma_1, \gamma_2, \ldots) \qquad (i = 1, 2, \ldots, n) \qquad (1)$$

Quantities such as the γ's have been called mental factors in recent psychological literature. However in view of the prospect of application of these ideas outside of psychology, and the conflicting usage attaching to the word "factor" in mathematics, it will be better simply to call the γ's *components* of the complex depicted by the tests.

Basic Idea

Reduce the dimensionality of a data set in which there is a large number (i.e., p is "large") of inter-related variables while retaining as much as possible the variation in the original set of variables

- The reduction is achieved by transforming the original variables to a new set of variables, "principal components", that are uncorrelated
- These principal components are ordered such that the first few retains most of the variation present in the data
- Goals/Objectives
  - Reduction and summary
  - Study the structure of covariance/correlation matrix

Some Applications

- Interpretation (by studying the structure of covariance/correlation matrix)
- Select a sub-set of the original variables, that are uncorrelated to each other, to be used in other multivariate procedures (e.g., multiple regression, classification)
- Detect outliers or clusters of multivariate observations

Multivariate Data

We display a multivariate data that contains n units on p variables using a matrix

$$\boldsymbol{X} = \begin{pmatrix} X_{1,1} & X_{2,1} & \cdots & X_{p,1} \\ X_{1,2} & X_{2,2} & \cdots & X_{p,2} \\ \vdots & \cdots & \ddots & \vdots \\ X_{1,n} & X_{2,n} & \cdots & X_{p,n} \end{pmatrix}$$

Summary Statistics

- Mean Vector: $\bar{\boldsymbol{X}} = (\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_p)^T$
- Covariance Matrix: $\Sigma = \{\sigma_{ij}\}_{i,j=1}^{p}$, where $\sigma_{ii} = \mathrm{Var}(X_i)$, $i = 1, \cdots, p$ and $\sigma_{ij} = \mathrm{Cov}(X_i, X_j), i \neq j$

Multivariate Data

We display a multivariate data that contains n units on p variables using a matrix

$$\boldsymbol{X} = \begin{pmatrix} X_{1,1} & X_{2,1} & \cdots & X_{p,1} \\ X_{1,2} & X_{2,2} & \cdots & X_{p,2} \\ \vdots & \cdots & \ddots & \vdots \\ X_{1,n} & X_{2,n} & \cdots & X_{p,n} \end{pmatrix}$$

Summary Statistics

- Mean Vector: $\bar{\boldsymbol{X}} = (\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_p)^T$
- Covariance Matrix: $\Sigma = \{\sigma_{ij}\}_{i,j=1}^{p}$, where $\sigma_{ii} = \mathrm{Var}(X_i)$, $i = 1, \cdots, p$ and $\sigma_{ij} = \mathrm{Cov}(X_i, X_j), i \neq j$

Next, we are going to discuss how to find **principal components**

Finding Principal Components

Principal Components (PCs) are uncorrelated **linear combinations** $\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_p$ determined sequentially, as follows:

1. The first PC is the linear combination $\tilde{X}_1 = \boldsymbol{c}_1^T \boldsymbol{X} = \sum_{i=1}^p c_{1i} X_i$ that maximize $\text{Var}(\tilde{X}_1)$ subject to $\boldsymbol{c}_1^T \boldsymbol{c}_1 = 1$

2. The second PC is the linear combination $\tilde{X}_2 = \boldsymbol{c}_2^T \boldsymbol{X} = \sum_{i=1}^p c_{2i} X_i$ that maximize $\text{Var}(\tilde{X}_2)$ subject to $\boldsymbol{c}_2^T \boldsymbol{c}_2 = 1$ and $\boldsymbol{c}_2^T \boldsymbol{c}_1 = 0$

$$\vdots$$

3. The p_{th} PC is the linear combination $\tilde{X}_p = \boldsymbol{c}_p^T \boldsymbol{X} = \sum_{i=1}^p c_{pi} X_i$ that maximize $\text{Var}(\tilde{X}_p)$ subject to $\boldsymbol{c}_p^T \boldsymbol{c}_p = 1$ and $\boldsymbol{c}_p^T \boldsymbol{c}_k = 0, \forall k < p$

Finding Principal Components by Decomposing Covariance Matrix

- Let Σ, the covariance matrix of $\boldsymbol{X}$, have eigenvalue-eigenvector pairs $(\lambda_i, \boldsymbol{e}_i)_{i=1}^p$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ Then, the k_{th} principal component is given by

$$\tilde{X}_k = \boldsymbol{e}_k^T \boldsymbol{X} = e_{k1} X_1 + e_{k2} X_2 + \cdots e_{kp} X_p$$

  $\Rightarrow$ we can perform a single matrix operation to get the coefficients to form all the PCs!

- Then,

$$\mathrm{Var}(\tilde{X}_i) = \lambda_i, \quad i = 1, \cdots, p$$

Moreover $\mathrm{Var}(\tilde{X}_1) \geq \mathrm{Var}(\tilde{X}_2) \geq \cdots \geq \mathrm{Var}(\tilde{X}_p) \geq 0$

$$\mathrm{Cov}(\tilde{X}_j, \tilde{X}_k) = 0, \quad \forall j \neq k$$

$\Rightarrow$ different PCs are uncorrelated with each other

PCA and Proportion of Variance Explained

- It can be shown that

$$\sum_{i=1}^{p} \mathrm{Var}(\tilde{X}_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \mathrm{Var}(X_i)$$

- The proportion of the total variance associated with the k_{th} principal component is given by

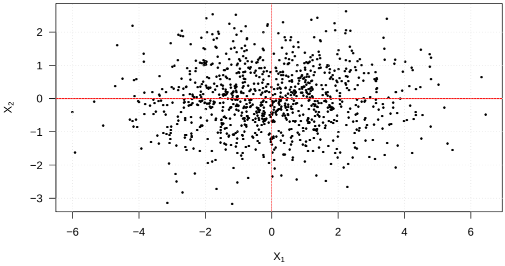
$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

- If a large proportion of the total population variance (say 80% or 90%) is explained by the first k PCs, then we can restrict attention to the first k PCs without much loss of information $\Rightarrow$ we achieve dimension reduction by considering $k < p$ uncorrelated components rather than the original p correlated variables

Toy Example 1

Suppose we have $\boldsymbol{X} = (X_1, X_2)^T$ where $X_1 \sim \mathrm{N}(0, 4)$, $X_2 \sim \mathrm{N}(0, 1)$ are independent

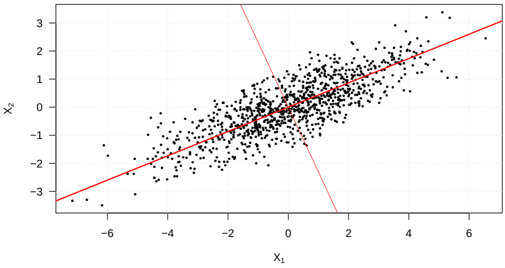
- Total variation $= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = 5$
- X_1 axis explains 80% of total variation
- X_2 axis explains the remaining 20% of total variation



Toy Example 2

Suppose we have $\boldsymbol{X} = (X_1, X_2)^T$ where $X_1 \sim \mathrm{N}(0,4)$, $X_2 \sim \mathrm{N}(0,1)$ and $\mathrm{Cor}(X_1, X_2) = 0.8$

- Total variation
  $= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = \mathrm{Var}(\tilde{X}_1) + \mathrm{Var}(\tilde{X}_2) = 5$
- $\tilde{X}_1 = .9175X_1 + .3975X_2$ explains 93.9% of total variation
- $\tilde{X}_2 = .3975X_1 - .9176X_2$ explains the remaining 6.1% of total variation



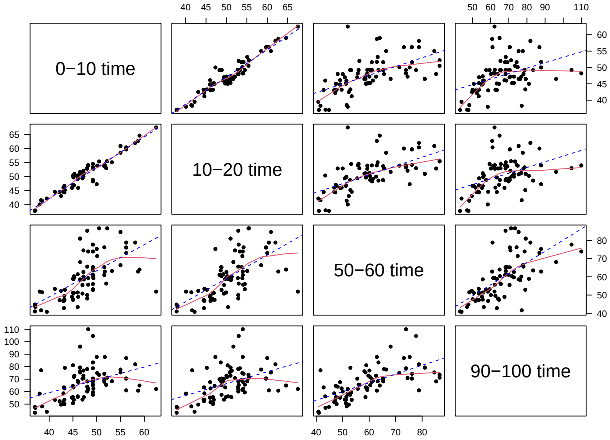
PCs of Standardized versus Original Variables

If we use standardized variables, i.e., $Z_j = \frac{X_j - \mu_j}{\sqrt{\sigma_{jj}}}$ $j = 1, \cdots, p$ ("z-scores"). Then we are going to work with the correlation matrix instead of the covariance matrix of $(X_1, \cdots, X_p)^{\mathrm{T}}$

- We can obtain PCs of standardized variables by applying spectral decomposition of the correlation matrix
- However, the PCs (and the proportion of variance explained) are, in general, different than those from original variables
- If units of p variables are comparable, covariance PCA may be more informative, if units of p variables are incomparable, correlation PCA may be more appropriate

Example: Men's 100k Road Race

The data consists of the times (in minutes) to complete successive 10k segments $(p = 10)$ of the race. There are 80 racers in total $(n = 80)$



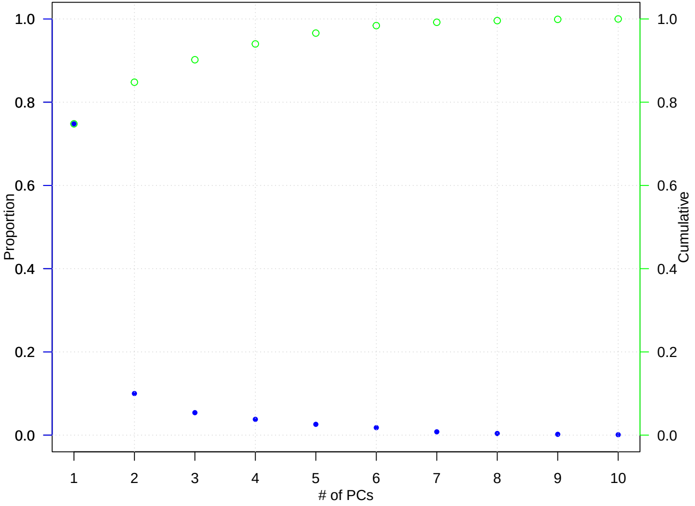
Eigenvalues of $\hat{\Sigma}$

	Eigenvalue	Proportion	Cumulative
PC1	735.77	0.75	0.75
PC2	98.47	0.10	0.85
PC3	53.27	0.05	0.90
PC4	37.30	0.04	0.94
PC5	26.04	0.03	0.97
PC6	17.25	0.02	0.98
PC7	8.03	0.01	0.99
PC8	4.25	0.00	1.00
PC9	2.40	0.00	1.00
PC10	1.29	0.00	1.00

Much of the total variance can be explained by the first three PCs

How Many Components to Retain?

A scree plot displays the variance explained by each component

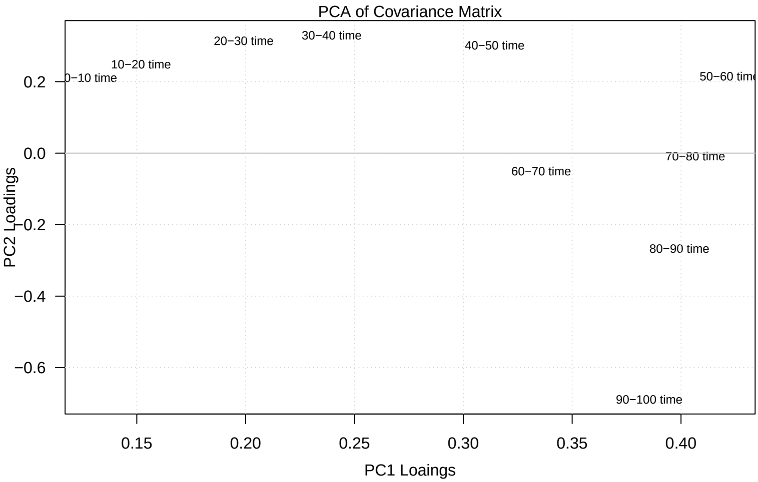


Men's 100k Road Race Component Weights

	Comp.1	Comp.2	Comp.3
0-10 time	0.13	0.21	0.36
10-20 time	0.15	0.25	0.42
20-30 time	0.20	0.31	0.34
30-40 time	0.24	0.33	0.20
40-50 time	0.31	0.30	-0.13
50-60 time	0.42	0.21	-0.22
60-70 time	0.34	-0.05	-0.19
70-80 time	0.41	-0.01	-0.54
80-90 time	0.40	-0.27	0.15
90-100 time	0.39	-0.69	0.35

What these numbers mean?

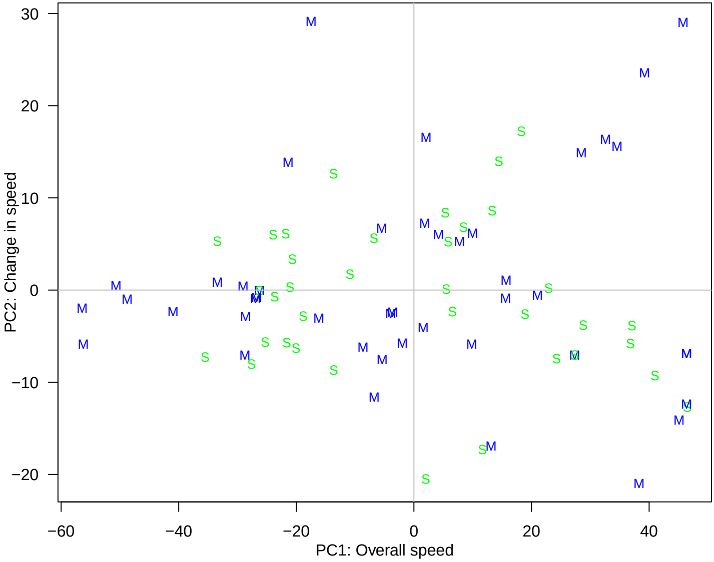
Visualizing the Weights to Gain Insight



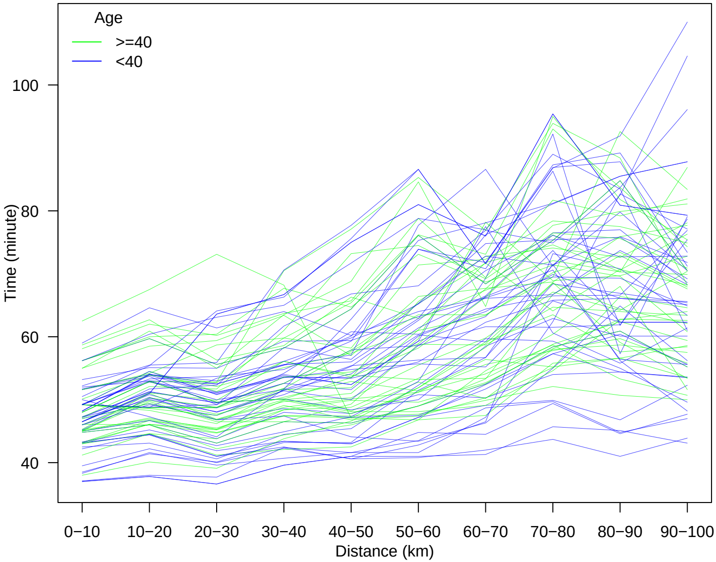
First component: overall speed
Second component: contrast long and short races

Looking for Patterns

Mature runners: Age < 40 (M); Senior runners: Age $>= 40$ (S)



Relating to Original Data: Profile Plot



Correlation PCA versus Covariance PCA