Chapter 5

Dimension Reduction and Extraction of Meaningful Factors

Section 5.1

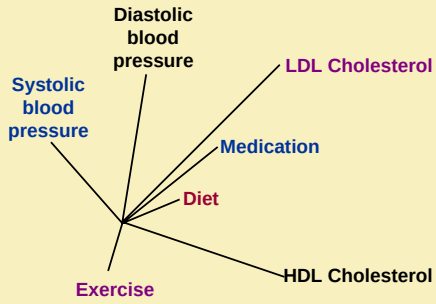
Principal Components Analysis

Objectives

- Explain the basic concepts for principal components analysis.
- Identify several strategies for selecting the number of components.
- Perform principal components analysis using the PRINCOMP procedure.

3

Too Many Variables



4

Solutions

- Eliminate some redundant variables.
  - May lose important information that was uniquely reflected in the eliminated variables.
- Create composite scores from variables (sum or average).
  - Lost variability among the variables
  - Multiple scale scores may still be collinear
- Create weighted linear combinations of variables while retaining most of the variability in the data.
  - Fewer variables; little or no lost variation
  - No collinear scales.

5

An Easy Choice

To retain most of the information in the data while reducing the number of variables you must deal with, try principal components analysis.

- Most of the variability in the original data can be retained.

but…

- Components may not be directly interpretable.

6

Principal Components Analysis

PCA

- is a dimension reduction method that creates variables called principal components
- creates as many components as there are input variables.

7

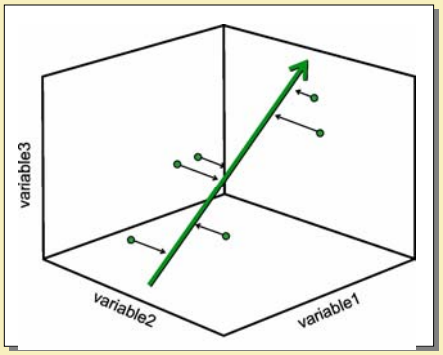
Principal Components

Principal components

- are weighted linear combinations of input variables
- are orthogonal to and independent of other components
- are generated so that the first component accounts for the most variation in the *x*s, followed by the second component, and so on.

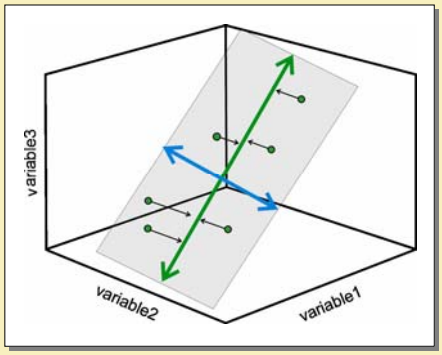
8

First Principal Component



9

Second Principal Component

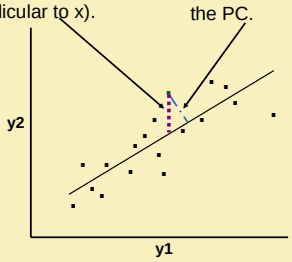


10

More on the Geometric Properties

Least squares regression minimizes the sum of squared *vertical* distances to the fitted line (perpendicular to x).

PCA minimizes the sum of the squared *perpendicular* distances to the axis of the PC.



11

Details of Principal Components

The *j* principal components provide a least-squares solution to the following model:

$$\mathbf{Y} = \mathbf{XB}$$

where

Y *n* by *p* matrix of scores on the components

X *n* by *j* matrix of centered observed variables

B *j* by *p* matrix of eigenvectors of the correlation or covariance matrix of the variables.

12

How Many Components?

- Scree plot of eigenvalues:

- Proportion of variance explained by each component: $\frac{\lambda_i}{\lambda_1+\lambda_2+..+\lambda_p}$ or $\frac{\lambda_i}{tr(\mathbf{R})}$
- Cumulative variance explained by components: $\frac{\lambda_1+\lambda_2+..+\lambda_k}{tr(\mathbf{R})}$
- Eigenvalue > 1

13

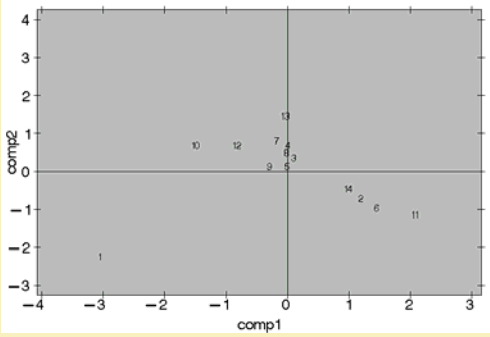
Principal Component Scores

Principal component scores can be created

- for each observation in the data set
- on each principal component
- using raw or the standardized weights.

14

Graphical Exploration of the PCs



15

Assumptions of PCA

- Random missingness
- Absence of outliers
- Singularity not a mathematical problem in PCA because matrices are not inverted.

16

Procedures That Can Perform PCA

- PRINCOMP
- PRINQUAL
- CORRESP
- PLS
- FACTOR.

17

The PRINCOMP Procedure

General form of the PRINCOMP procedure:

```
PROC PRINCOMP <options>;
     VAR variables;
RUN;
```

18

The FACTOR Procedure

General form of the FACTOR procedure:

```
PROC FACTOR options;
     VAR variables;
RUN;
```

19