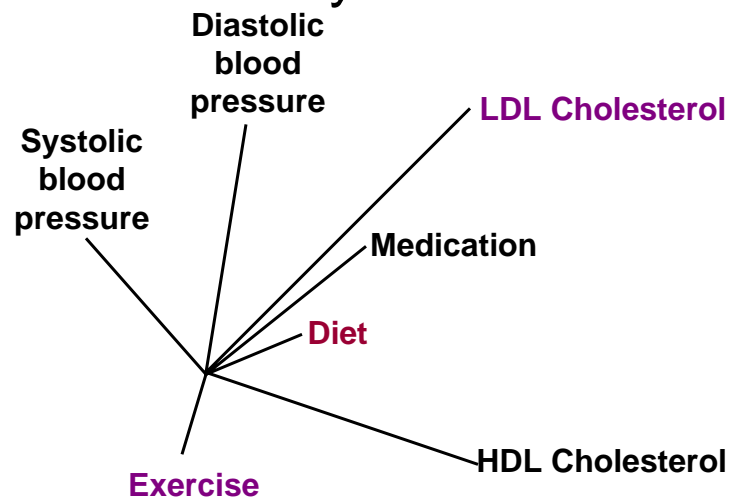


Principal Components Analysis

Too Many Variables



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.

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.

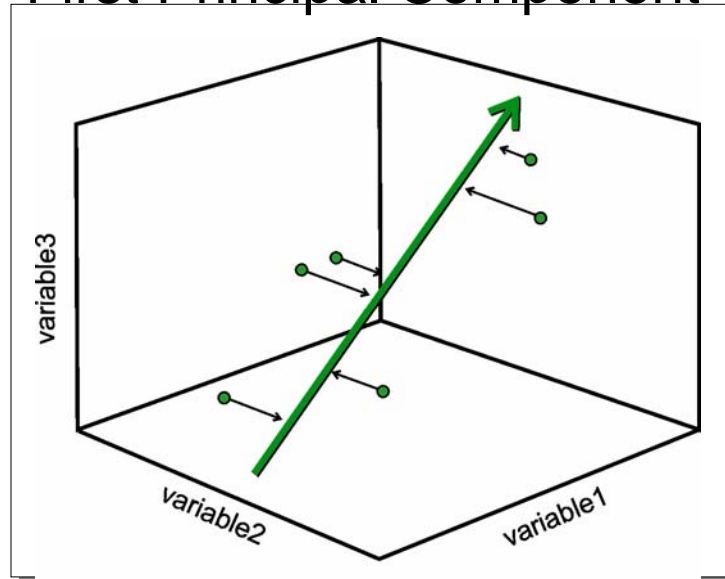
Principal Components Analysis

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

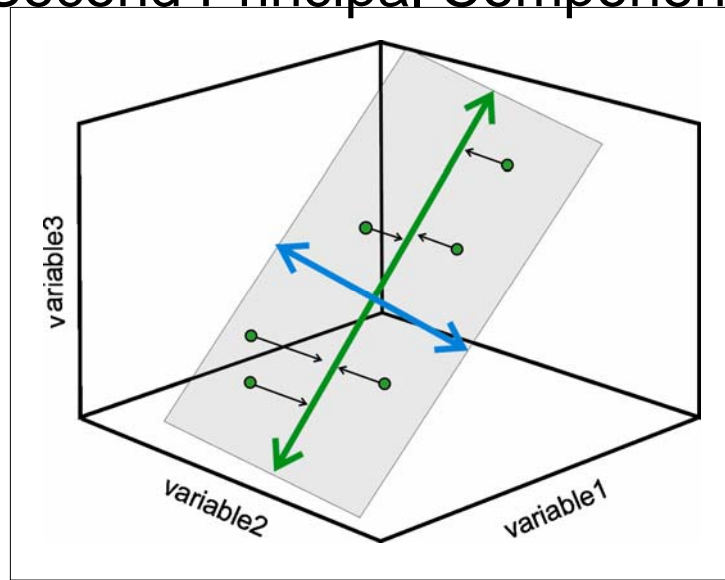
Principal Components

- *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.

First Principal Component



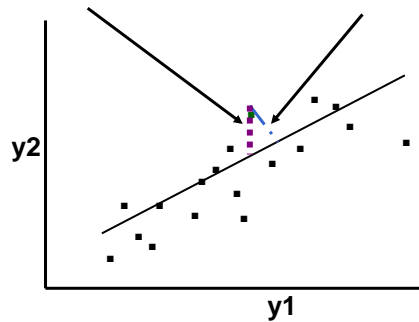
Second Principal Component



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.



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.

Uses of PCA

- Data Screening
- Visual Clustering of observations
- Reduction of Dimensionality

PCA on Σ

Let $\mathbf{x} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$

The spectral decomposition of $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}'$$

where

$$\mathbf{A}\mathbf{A}' = \mathbf{I}$$

PCA on Σ

Let $\mathbf{y} = \mathbf{A}'(\mathbf{x} - \boldsymbol{\mu})$

Then

$\mathbf{y}_1 = \mathbf{a}'_1(\mathbf{x} - \boldsymbol{\mu})$ is the first principal component

$\mathbf{y}_2 = \mathbf{a}'_2(\mathbf{x} - \boldsymbol{\mu})$ is the second principal component

First Principal Component

= linear combination $\mathbf{a}'_1(\mathbf{x} - \boldsymbol{\mu})$

that maximizes $\text{Var}[\mathbf{a}'_1(\mathbf{x} - \boldsymbol{\mu})]$

subject to $\mathbf{a}'_1\mathbf{a}_1 = 1$

Second Principal Component

= linear combination $\mathbf{a}'_2(\mathbf{x} - \boldsymbol{\mu})$

that maximizes $\text{Var}[\mathbf{a}'_2(\mathbf{x} - \boldsymbol{\mu})]$

subject to $\mathbf{a}'_2\mathbf{a}_2 = 1$

and

$$\text{Cov}[\mathbf{a}'_1(\mathbf{x} - \boldsymbol{\mu}), \mathbf{a}'_2(\mathbf{x} - \boldsymbol{\mu})] = 0$$

i^{th} Principal Component

= linear combination $\mathbf{a}'_i(\mathbf{x} - \boldsymbol{\mu})$

that maximizes $\text{Var}[\mathbf{a}'_i(\mathbf{x} - \boldsymbol{\mu})]$

subject to $\mathbf{a}'_i\mathbf{a}_i = 1$

and

$$\text{Cov}[\mathbf{a}'_i(\mathbf{x} - \boldsymbol{\mu}), \mathbf{a}'_j(\mathbf{x} - \boldsymbol{\mu})] = 0$$

for $j < i$

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.

Principal Component Scores

$$\text{Let } \mathbf{X}_{N \times p} = \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_N \end{bmatrix}$$

Principal Component Scores

Then

$$\mathbf{Y}_{N \times p} = \left\{ \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_N \end{bmatrix} - \begin{bmatrix} \mathbf{u}' \\ \mathbf{u}' \\ \vdots \\ \mathbf{u}' \end{bmatrix} \right\} \mathbf{A} = (\mathbf{X} - \mathbf{1}\boldsymbol{\mu}') \mathbf{A}$$

defines the principal component scores

Principal Component Loading Vectors

$$\mathbf{c}_j = \sqrt{\lambda_j} \mathbf{a}_j$$

Allows comparisons among all the elements in all the vectors \mathbf{c}'_j

The i^{th} element in the j^{th} component loading vector gives the covariance between the i^{th} original variable and the j^{th} principal component.

Determining the Number of Principal Components

1. proportion of variance accounted for by each principal components:

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + \dots + \lambda_p} = \frac{\lambda_j}{\text{tr}(\mathbf{\Sigma})}$$

or

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + \dots + \lambda_p} = \frac{\lambda_j}{\text{tr}(\mathbf{P})}$$

Determining the Number of Principal Components

2. % of variance accounted for by the first k principal components:

$$\frac{\lambda_1 + \lambda_2 + \dots + \lambda_k}{\text{tr}(\mathbf{\Sigma})}$$

or

$$\frac{\lambda_1 + \lambda_2 + \dots + \lambda_k}{\text{tr}(\mathbf{P})}$$

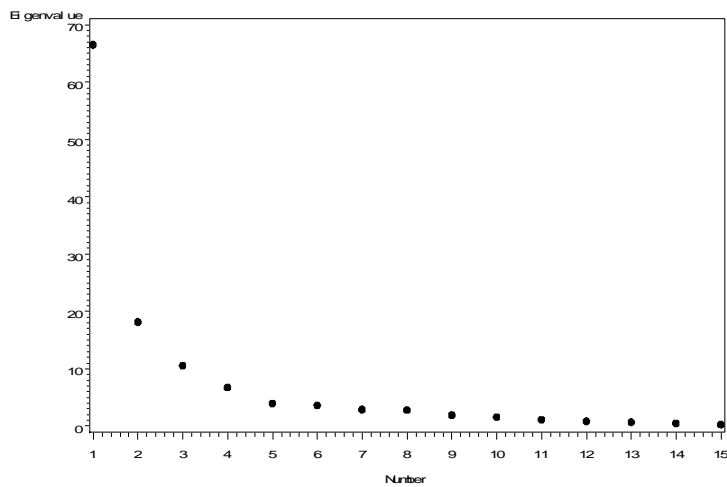
Determining the Number of Principal Components

3. Eigenvalue > 1

- When doing a PCA on the correlation matrix R , the sum of the eigenvalues is equal to the number of variables in the analysis
- Principal components with eigenvalues greater than 1 explain more of the variance than any single variable in the analysis.

Determining the Number of Principal Components

4. Scree Plot



Appropriateness of PCA on Σ

1. All variables measured on the same units
2. All variables have approximately equal variance

PCA on P

- If either of the above conditions is not true, it is often more appropriate to determine the principal components from the correlation matrix P , or, in practice, the sample estimate R .

Assumptions of PCA

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

Testing Independence of Variables

$$H_0 : \mathbf{P} = \mathbf{I}$$

$$H_1 : \mathbf{P} \neq \mathbf{I}$$

The Likelihood Ratio test statistic is:

$$\left[(N-1) - \frac{(2p+5)}{6} \right] \log |\mathbf{R}| \sim \chi_{p(p-1)/2}^2$$

SAS Procedures That Can Perform PCA

- PRINCOMP
- PRINQUAL
- CORRESP
- PLS
- FACTOR