Knowledge Discovery and Data Mining

Unit # 16

Source

- Applied Multivariate Statistical Analysis by Richard Johnson and Dean Wichern, 2002
- Using Multivariate Statistics by Barbara Tabachnick and Linda Fidell, 1996

Introduction

- Principal comonent analysis (PCA) and factor analysis (FA) are statistical techniques applied to a single set of variables where the researcher is interested in discovering which variables in the set form coherent subsets that are relatively independent of one another.
- Variables that are correlated with one another but largely independent of other subsets of variables are combined into factors.
- Factors are thought to reflect underlying processes that have created the correlations among variables.

Purpose of PCA and FA

- The specific goals of PCA or FA are to
  - summarize patterns of correlations among observed variables,
  - to reduce a large number of observed variables to a smaller number of factors,
  - to provide an operational definition for an underlying process by using observed variables or
  - to test a theory about the nature of underlying process

Fundamental Steps

- Steps in PCA or FA include
  - Selecting and measuring a set of variables
  - Preparing the correlation matrix
  - Extracting a set of factors from the correlation matrix
  - Determining the number of factors
  - interpreting the results
- Although there are relevant statistical considerations to most of these steps, an important test of the analysis is its interpretability.

Application

- PCA is a useful statistical technique that has found application in fields such as face recognition and image compression, and is a common technique for finding patterns in data of high dimension.

Limitation

- One of the problems with PCA and FA is that there is no criterion variable against which to test the solution.
- In regression analysis, for instance, the dependent variable (DV) is a criterion and the correlation between observed and predicted DV scores serves as a test of the solution
- In classification, the solution is judged by how well it predicts group membership.
- But in PCA or FA there is no external criterion such as group membership against which to test the solution.

Practical Issues

- Because FA and PCA are exquisitely sensitive to the sizes of correlations, it is critical that honest, reliable correlations be employed.
- Sensitivity to outlying cases, problems created by missing data, and degradation of correlations between poorly distributed variables all plague FA and PCA.

Normality

- As long as PCA and FA are used as convenient ways to summarize the relationships in a large set of observed variables, assumptions regarding the distributions of variables are not in force.
- If variables are normally distributed, the solution is enhanced. To the extent that normality fails, the solution is degraded but may still be worthwhile.
- However, multivariate normality is assumed when statistical inference is used to determine the number of factors. Multivariate normality is the assumption that all variables, and all linear combinations of variables, are normally distributed.

Variance and Covariance

- Standard deviation and variance only operate on one dimension.
- However, it is useful to have a similar measure to find out how much the dimensions vary from the mean *with respect to each other.*
- Covariance is such a measure. Covariance is always measured *between* 2 dimensions.
- If you calculate the covariance between one dimension and *itself, you get the* variance.
- So, if you had a 3-dimensional data set (x, y, z), then you could measure the covariance between the x an y dimensions, the x and z dimensions, and the y and z dimensions.

Variance and Covariance (Cont'd)

$$var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{(n-1)}$$

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

Eigen Values and Eigen Vectors

- Let *A* be a square matrix. A non-zero vector *C* is called an **eigenvector** of *A* if and only if there exists a number (real or complex) λ such that

$$AC = \lambda C$$

- If such a number λ exists, it is called an **eigenvalue** of *A*. The vector *C* is called eigenvector associated to the eigenvalue .

Eigen Vectors

- Eigenvectors can only be found for square matrices.
- And, not every square matrix has eigenvectors.
- And, given an nxn matrix that does have eigenvectors, there are n of them.
- For example, given a 3x3 matrix, there are 3 eigenvectors.

Eigen Value Computation

- When a transformation is represented by a square matrix A, the eigen value equation can be expressed as $Ax - \lambda x = 0$
- Where I is the identify matrix. The can be rearranged to $(A - \lambda I)x = 0$
- If there exists an inverse $(A - \lambda I)^{-1}$ then both sides can be left multiplied by the inverse to obtain the trivial solutions: $x = 0$. Thus we require there to be no inverse by assuming from linear algebra that the determinants equals zero:
- $\det(A - \lambda I) = 0$
- To compute eigen vectors, solve for $Ax = \lambda x$ for all values of λ.

Example (Source: Wikipedia)

For the matrix A

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

the vector

$$\mathbf{x} = \begin{bmatrix} 3 \\ -3 \end{bmatrix}$$

is an eigenvector with eigenvalue 1. Indeed,

$$A\mathbf{x} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ -3 \end{bmatrix} = \begin{bmatrix} 2 \cdot 3 + 1 \cdot (-3) \\ 1 \cdot 3 + 2 \cdot (-3) \end{bmatrix} = \begin{bmatrix} 3 \\ -3 \end{bmatrix} = 1 \cdot \begin{bmatrix} 3 \\ -3 \end{bmatrix}.$$

On the other hand the vector

$$\mathbf{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

is *not* an eigenvector, since

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \cdot 0 + 1 \cdot 1 \\ 1 \cdot 0 + 2 \cdot 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

and this vector is not a multiple of the original vector $\mathbf{x}$.

Example

Exercises

For the following square matrix:

$$\begin{pmatrix} 3 & 0 & 1 \\ -4 & 1 & 2 \\ -6 & 0 & -2 \end{pmatrix}$$

Decide which, if any, of the following vectors are eigenvectors of that matrix and give the corresponding eigenvalue.

$$\begin{pmatrix} 2 \\ 2 \\ -1 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \\ 3 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$$

Working of PCA

- The original data are projected onto a much smaller space, resulting in dimensionality reduction.
- Unlike attribute subset selection, which reduces the attribute set size by retaining a subset of the initial set of attributes, PCA "combines" the essence of attributes by creating an alternative, smaller set of variables.
- The initial data can then be projected onto this smaller set.

Steps 1 and 2

- Step 1:
  - The input data are normalized, so that each attribute falls within the same range. This step helps ensure that attributes with large domains will not dominate attributes with smaller domains.
- Step 2:
  - PCA computes *k orthonormal vectors that provide a basis for the normalized input* data.
  - These are unit vectors that each point in a direction perpendicular to the others.
  - These vectors are referred to as the *principal components.*
  - *The input data are a linear* combination of the principal components.

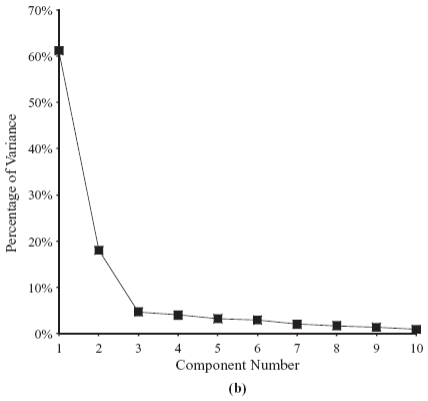
Steps 3 and 4

- Step 3:
  - The principal components are sorted in order of decreasing "significance" or strength.
  - The principal components essentially serve as a new set of axes for the data, providing important information about variance.
  - That is, the sorted axes are such that the first axis shows the most variance among the data, the second axis shows the next highest variance, and so on.
- Step 4:
  - Because the components are sorted according to decreasing order of "significance," the size of the data can be reduced by eliminating the weaker components, that is, those with low variance.
  - Using the strongest principal components, it should be possible to reconstruct a good approximation of the original data.

Example (Source: Witten et al.)

Axis	Variance	Cumulative
1	61.2%	61.2%
2	18.0%	79.2%
3	4.7%	83.9%
4	4.0%	87.9%
5	3.2%	91.1%
6	2.9%	94.0%
7	2.0%	96.0%
8	1.7%	97.7%
9	1.4%	99.1%
10	0.9%	100.0%

(a)



(b)

FIGURE 7.5

Principal components transform of a dataset: (a) variance of each component and (b) variance plot.

Example

- Analyze iris data using R
  - X <- cbind(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width)
- Computer covariance
  - X_Cov <- cov(X)
- Compute eigen values and vectors
  - X_Eig <- eigen(X_Cov)
- We can perform PCA directly as well
  - X_PCA <- princomp(X, cor=FALSE)
  - summary(X_PCA)
  - loadings(X_PCA)
  - plot(X_PCA, type="lines")
  - Y <- X_PCA$scores
  - cor(X, Y)

KNIME Demo