Lecture 5 :
Quantitative multivariate data mining
-
Dimensionality Reduction
-
Linear Discriminant Analysis

3 décembre 2012

Introduction

Summary

When each data is described by several random variables the dimensionality reduction methods are often used that allow the data visualization and the structure detection.

Suppose thet the data ar classified (labels are known). Here we want to view the best differences between groups of data. We dispose information on the membership of each data to group (eg. men and women) and we search for a representation in two or even three dimensions that separates the best possible the groups.

Goals

We are looking for new variables (discriminating variables), corresponding to the axes of the representation space, which best separate (in projection) the k groups (or classes) of observations.

Principle

The linear discriminant analysis is a method of data analysis and dimensionality reduction which is very close to the the PCA. The only difference is that we want to maximize the separation of classes along the projection of axes, instead of the variance of the data.

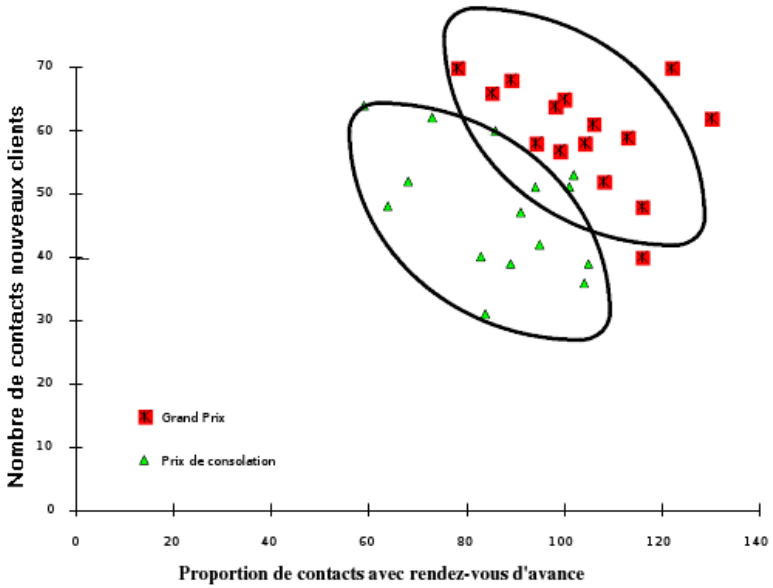
Principle

General Principle :

1. Compute the total covariance between variables, but also the intragroup covariance and the covariance between groups.
2. These indices are used to create new variables :
   - Creates new variables (same numbre as old variables).
   - Each new variable is a linear combination of all the original variables.
   - These new variables must be totally independent of each other (uncorrelated).
3. Select among the new variables as those representing the distribution of data in different classes and eliminates the others.

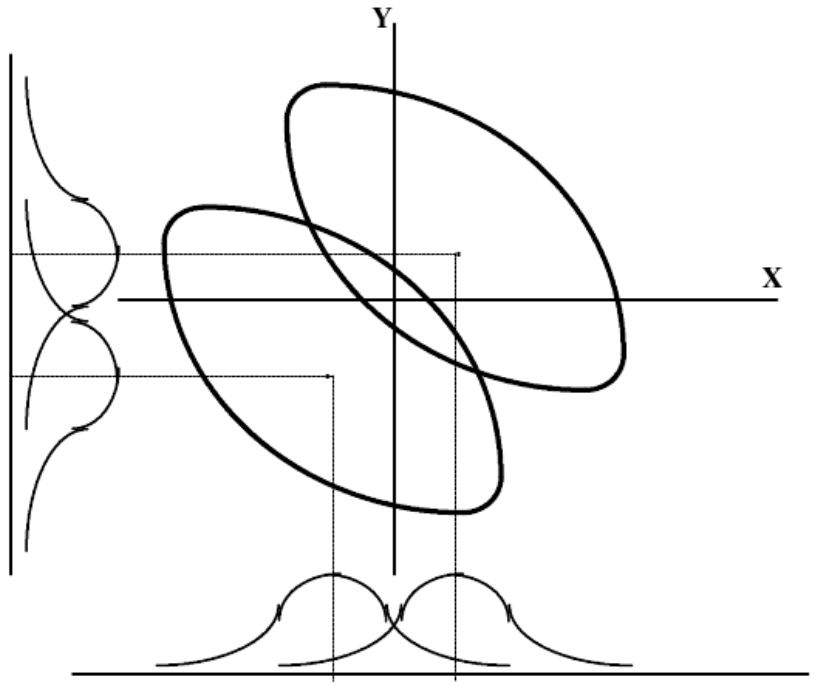
Principle

We can represent the distribution of data classes with a minimum of loss :



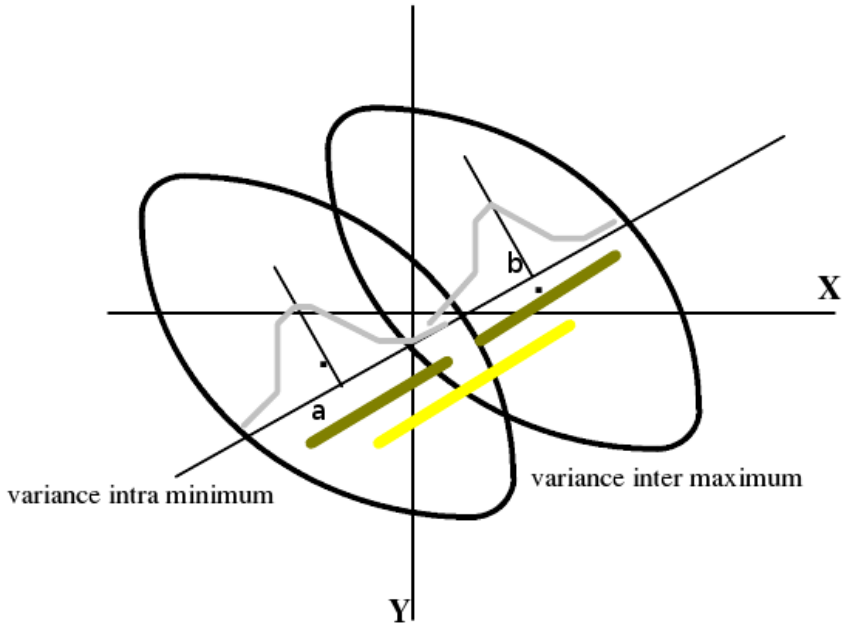
Principle

We can represent the distribution of data classes with a minimum of loss :



Principle

We can represent the distribution of data classes with a minimum of loss :



Calculation of covariances

total covariance

Let m_{tot} be the mean of all N objects M and let $\bar{M}$ - the centered matrix on m_{tot}. The total covariance matrix is easily computed by a matrix product :

Total covariance matrix

$$C_{tot} = \frac{1}{N-1} \cdot {}^t\bar{M} \cdot \bar{M}$$

This matrix represents the dispersion of all the sample data in the representation space.

Intra-class covariance

Assume that there are k classes. Let m_i be the mean of the data M_i of the class i and let $\bar{M}_i$ be the centered matrix on m_i. The within-class covariance matrix is calculated as follows :

The within-class covariance matrix (intra-class)

$$C_{intra} = \frac{1}{N-1} \cdot \sum_{i=1}^{k} {}^t\bar{M}_i \cdot \bar{M}_i$$

This matrix represents the dispersion of the data within groups.

Inter-class covariance

Finnally, let N_i be the number of objects which belongs to the class i. The inter-class covariance matrix is calculated as follows :

The inter-class covariance :

$$C_{inter} = \frac{1}{N-1} \cdot \sum_{i=1}^{k} N_i \cdot {}^t(m_i - m) \cdot (m_i - m)$$

This matrix represents the dispersion of the groups in the representation space.

Fundamental relationship

The covariance is the sum of the inter-class covariance and within-class covariance

Fundamental relationship :

$$C_{tot} = C_{inter} + C_{intra}$$

Classification of new objects

It is possible to determine automatically the class K of a new presented object x using the within-class covariance matrix :

Automatic classification of a new observation :

$$K = \underset{i}{Argmin}\left((x - m_i) \cdot (C_{intra})^{-1} \cdot {}^{t}(x - m_i)\right)$$

Calculation of new variables

Summary

Now that we know the covariance relationship between variables we seek to create, by linear combinations, a new variable represented by a single axis u_1 i.e. the projection of the data on u_1 has a maximum inter-class variance and a minimum intra-class variance. Then seek a second axis u_2 independent of u_1 (i.e. orthogonal to u_1) that best explains the remaining variance, and so on ...

New variables

We seek for the vector u of norm 1 such that the projection of the data points on u has a maximum between-class (inter-class) variance for a minimum within-class (intra-class) variance.

The projection of the sample X on u is noted as follows :

$$\pi_u(M) = M \cdot u$$

Total variances, inter et intra-class of $\pi_u(M)$ are equals to :

$$V^{\pi}_{tot} = {}^t u \cdot C_{tot} \cdot u$$

$$V^{\pi}_{inter} = {}^t u \cdot C_{inter} \cdot u$$

$$V^{\pi}_{intra} = {}^t u \cdot C_{intra} \cdot u$$

New variables

We seek to maximize the inter-class variance and to minimize the within-class variance. Since $C_{tot} = C_{inter} + C_{intra}$, this amounts to maximize the proportion of C_{inter} on C_{tot}. We will therefore seek to maximize the projection of $\frac{C_{inter}}{C_{tot}}$ on u, ie :

We seek to maximize :

$$J = \frac{V^{\pi}_{inter}}{V^{\pi}_{tot}} = \frac{{}^{t}u \cdot C_{inter} \cdot u}{{}^{t}u \cdot C_{tot} \cdot u}$$

New variables

Maximize J means to solve :

$$(C_{tot})^{-1} \cdot C_{inter} \cdot u = J \cdot u$$

By definition, the solution of this equation is the set of couples : eigenvectors / eigenvalues of $(C_{tot})^{-1} \cdot C_{inter}$.

- The bigest eignevalue λ_1 is the maximum value of J.
- The associated eigenvector represents the projection axis u_1 along which J is maximal.

New variables

The above reasoning has allowed us to write the vector that best explains the dispersion of the group is the first eigenvector. Similarly, the second vector that best explains the remaining dispersion of groups (not explained by the first axis) is the second eigenvector, etc..

We saw also that the dispersion (also called discriminatory power) J of groups explained by the n^{th} eigenvector is λ_n.

Dimensionality reduction

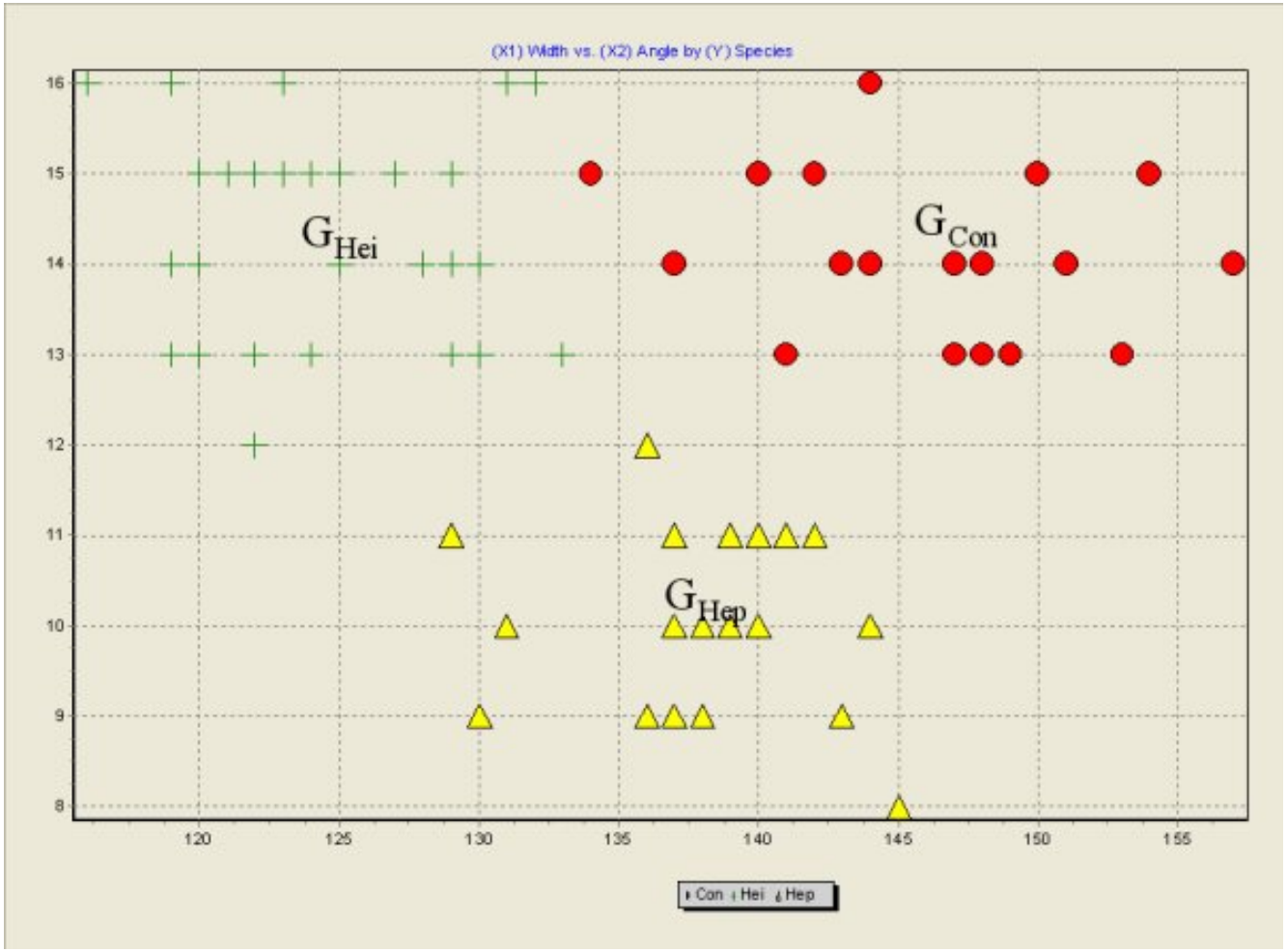
To reduce the dimensionality of the dataset :

1. Choose X the number of dimensions to keep.
2. Select X eigenvectors $u_1 ... u_x$ associated with X larger eigenvalues.
3. Project the data onto the new axes (also called discriminant axes) for a description of these data using new variables :

$$\pi_u(M) = M \cdot u$$

4. If the number of dimensions is ≤ 3, we can represent data in the new space.

Dimensionality reduction



Dimensionality reduction

Another possibility :

1. We choose P the proportion of information that you want to keep.
2. Select the minimum number X of eigenvector $u_1 ... u_x$ associated to X eigenvalues $\lambda_1 ... \lambda_x$ i.e. :

$$\frac{\sum_{i=1}^{x} \lambda_i}{\sum \lambda} \geq P$$

3. Project the data on new axes.
4. If the number of dimensions is ≤ 3, we can represent data in the new space.

Analysis of axes

Correlations between old and new variables

We know that the new variables are linear combinations of the old ones. It is therefore interesting to see the correlation between the old and new variables to interpret the results.

Correlations between old and new variables

Correlations between an old and a new variable variable :

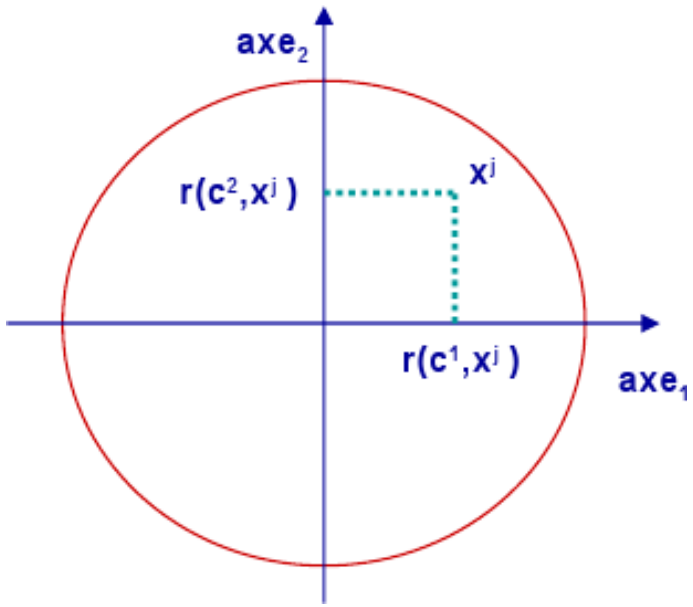
$$r(M_i, \pi_j) = \frac{\frac{1}{N-1} \cdot {}^t\overline{M_i} \cdot \overline{\pi_j}}{\sigma_i \sigma_j}$$

With $\overline{M_i}$ and i^{eme} columns of the matrix $\overline{M}$ containing centered data, $\overline{\pi_j}$ represents the projection of $\overline{M}$ on the new axis j. This expression follows directly from the formula for the correlation coefficient. Unlike PCA, here we can **not** use λ_j instead of the variance of the data projection on the new axis u_j, since in this case λ_j is a group separation value.

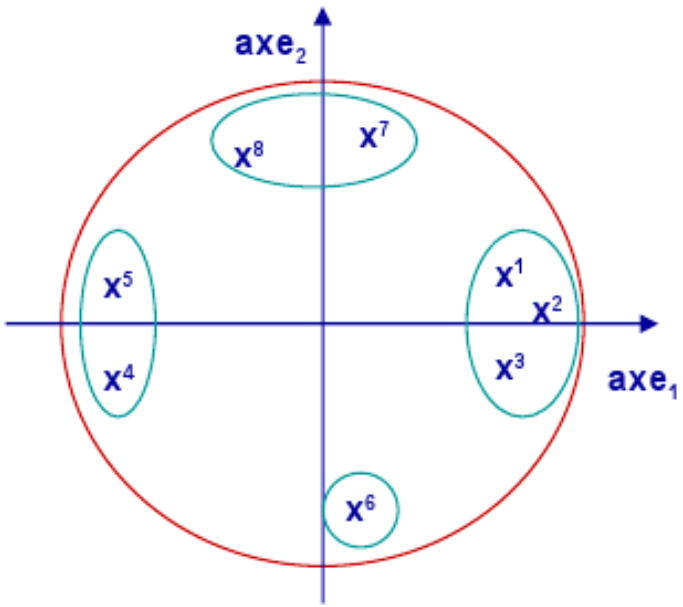
Visualization

Now that we know the correlation between the old and new variables, we can simply visualize a circle of “correlations”. This allows both to characterize new variables (which allows us to visually interpret the data) and visualize the correlation between the old variables.

Visualization



Visualization



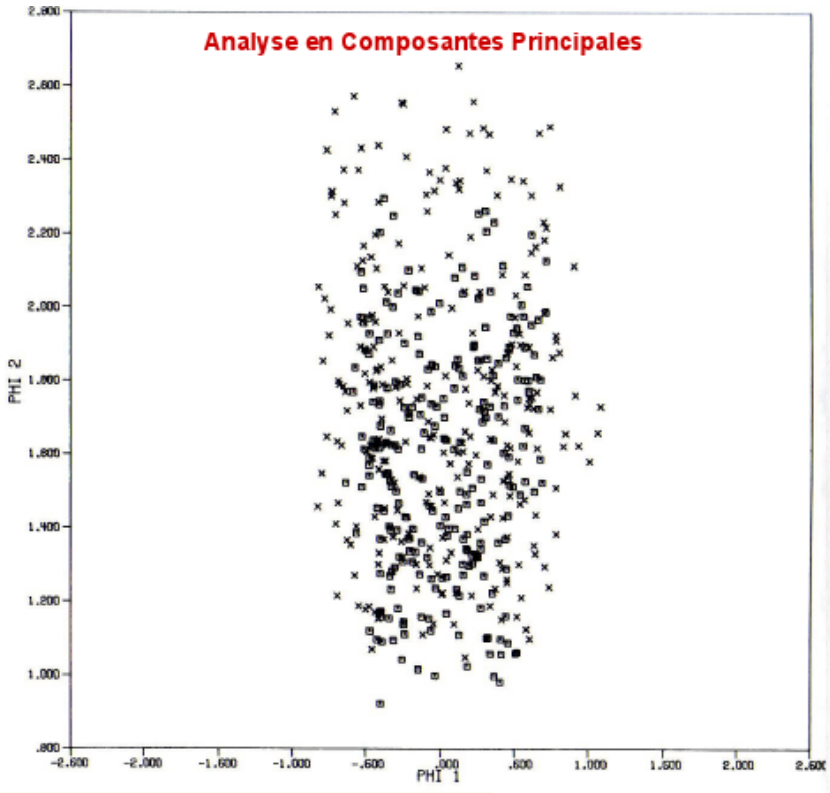
Conclusion

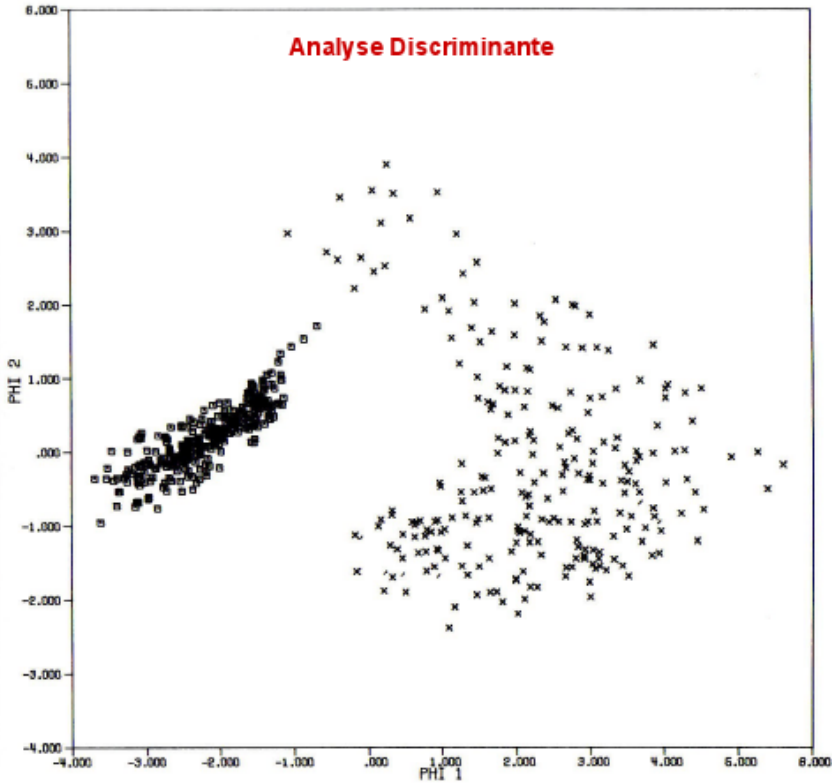
Conclusion

Information that can be obtained with Linear Discriminant Analysis :

1. Reduced dimensions allowing effective visalization (scatter) of the data groups.
2. The power of accurate discrimination of groups contained in each new variable.
3. Correlations between the old and new variables.
4. Visualization of correlations between the old variables.

PCA vs LDA


Analyse en Composantes Principales
PHI 2
PHI 1


Analyse Discriminante
PHI 2
PHI 1