Descriptive Discriminant Analysis
(Canonical Discriminant Analysis)

Multivariate characterization of differences between groups

Ricco RAKOTOMALALA
Outline

1. Problem statement
2. Determination of the latent variables (dimensions)
3. Reading the results
4. A case study
5. Classification of a new instance
7. Conclusion
8. References
Issues

From a set of quantitative variables, how to compute a new representation space (dimensions) which enables to highlight the differences between groups of individuals
Descriptive Discriminant Analysis (DDA) - Goal

A population is subdivided in $K$ groups (using a categorical variable, a label); the instances are described by $J$ continuous descriptors.

E.g. Bordeaux wine (Tenenhaus, 2006; page 353). The rows of the dataset correspond to the year of production (1924 to 1957)

<table>
<thead>
<tr>
<th>Annee</th>
<th>Temperature</th>
<th>Soleil</th>
<th>Chaleur</th>
<th>Pluie</th>
<th>Qualite</th>
</tr>
</thead>
<tbody>
<tr>
<td>1924</td>
<td>3064</td>
<td>1201</td>
<td>10</td>
<td>361</td>
<td>medium</td>
</tr>
<tr>
<td>1925</td>
<td>3000</td>
<td>1053</td>
<td>11</td>
<td>338</td>
<td>bad</td>
</tr>
<tr>
<td>1926</td>
<td>3155</td>
<td>1133</td>
<td>19</td>
<td>393</td>
<td>medium</td>
</tr>
<tr>
<td>1927</td>
<td>3085</td>
<td>970</td>
<td>4</td>
<td>467</td>
<td>bad</td>
</tr>
<tr>
<td>1928</td>
<td>3245</td>
<td>1258</td>
<td>36</td>
<td>294</td>
<td>good</td>
</tr>
<tr>
<td>1929</td>
<td>3267</td>
<td>1386</td>
<td>35</td>
<td>225</td>
<td>good</td>
</tr>
</tbody>
</table>

**Goal(s):**

1. Descriptive (explanation): highlighting the characteristics which enable to explain the differences between groups → main objective in our context

2. Predictive (classification): assign a group to an unseen instance → secondary objective in our context (but this is the main objective in the predictive discriminant analysis [PDA] context)
Descriptive Discriminant Analysis - Approach

**Aim:** Determining the most parsimonious way to explain the differences between groups by computing a set of orthogonal linear combinations (canonical variables, factors) from the original descriptors. *Canonical Discriminant Analysis.*

The conditional centroids must be as widely separated as possible on the factors.

\[
\sum_i (z_i - \bar{z})^2 = \sum_k n_k (\bar{z}_k - \bar{z})^2 + \sum_k \sum_i (z_{ik} - \bar{z}_k)^2
\]

\[v = b + w\]

*Total* (variation) = *Between* class (variation) + *Within* class (variation)
Maximizing a measure of the class separability: the correlation ratio.

\[ \eta_{z,y}^2 = \frac{b}{v} \quad \text{with} \quad 0 \leq \eta_{z,y}^2 \leq 1 \]

1 \(\rightarrow\) Perfect discrimination. All the points related to a group are confounded to the corresponding centroid \((W = 0)\)

0 \(\rightarrow\) Impossible discrimination. All the centroids are confounded \((B = 0)\)

Determining the coefficients (canonical coefficients) \((a_1,a_2)\) which maximize the correlation ratio

Maximum number of “dimensions” (factors):
\[ M = \min(J, K-1) \]

The factors are uncorrelated

A factor takes into account the differences not explained by the preceding factors

The correlation ratio measures the class separability
Solution

How to compute the canonical variables that summarize the between-class variation
Descriptive Discriminant Analysis
Mathematical formulation

\[ a = \begin{pmatrix} a_1 \\ \vdots \\ a_j \end{pmatrix} \]

« \( a \) » is the vector of coefficients which enables to define the canonical variable \( Z \) i.e.
\[ z = a_1(x_i - \bar{x}_l) + \cdots + a_j(x_j - \bar{x}_j) \]

---

**Huyghens’ theorem \( \Rightarrow V = B + W \)**

Total covariance matrix

\[ V \rightarrow v_{lc} = \frac{1}{n} \sum_i \left( x_{il} - \bar{x}_l \right) \left( x_{ic} - \bar{x}_c \right) \]

Within groups covariance matrix

\[ W \rightarrow w_{lc} = \frac{1}{n} \sum_k \sum_{i:y_i = k} \left( x_{il,k} - \bar{x}_{l,k} \right) \left( x_{ic,k} - \bar{x}_{c,k} \right) \]

Between groups covariance matrix

\[ B \rightarrow b_{lc} = \sum_k \frac{n_k}{n} \left( \bar{x}_{l,k} - \bar{x}_l \right) \left( \bar{x}_{c,k} - \bar{x}_c \right) \]

The aim of DDA is to calculate the coefficients of the canonical variable which maximizes the correlation ratio

\[ \max_a \frac{a' Ba}{a' Va} \equiv \max_a \eta_{z,y}^2 \]

---

**Total sum of squares**

\[ TSS = a' Va \quad \text{[ignoring a multiplication factor (1/n)]} \]

\[ RSS = a' Wa \]

\[ ESS = a' Ba \]
Descriptive Discriminant Analysis

Solution

\[
\max_a \frac{a' B a}{a' V a}
\]

is equivalent to

\[
\max_a a' B a
\]

Under the constraint \( a' V a = 1 \) \( (\text{“} a \text{” is a unit vector}) \)

Solution: using the Lagrange function \( (\lambda \text{ is the Lagrange multiplier}) \)

\[
L(a) = a' B a - \lambda (a' V a - 1)
\]

\[
\frac{\partial L(a)}{\partial a} = 0 \Rightarrow B a = \lambda V a
\]

\[\Rightarrow V^{-1} B a = \lambda a\]

\( \lambda \) is the first eigenvalue of \( V^{-1} B \)

“\( a \)” is the corresponding eigenvector

The successive canonical variables are obtained from the eigenvalues and the eigenvectors of \( V^{-1} B \).

The number of non-zero eigenvalue is \( M = \min(K-1, J) \) i.e. \( M \) canonical variables

\[
\lambda = \eta^2 \quad \text{The eigenvalue is equal to the square of the correlation ratio} \ (0 \leq \lambda \leq 1)
\]

\[
\eta = \sqrt{\lambda} \quad \text{is the canonical correlation}\]
Discriminant descriptive analysis
Bordeaux wine (X1 : Temperature and X2 : Sun)

\[
Z_{i2} = -0.0092(x_{i1} - \bar{x}_1) + 0.0105(x_{i2} - \bar{x}_2)
\]

\[
\eta_2 = \sqrt{0.051} = 0.225
\]

The differences between the centroids are lesser on this factor.

\[
Z_{i1} = 0.0075(x_{i1} - \bar{x}_1) + 0.0075(x_{i2} - \bar{x}_2)
\]

\[
\eta_1 = \sqrt{0.726} = 0.852
\]

The differences between the centroids are high on this factor.

(2.91; -2.22): the coordinates of the individuals in the new representation space are called “factor scores” (SAS, SPSS, R...)
Discriminant descriptive analysis
Alternative solution – English-speaking tools and references

Since \( V = B + W \), we can formulate the problem in other way:

\[
\max_a \frac{a' Ba}{a' Wa} \quad \text{is equivalent to} \quad \max_a \frac{a' Ba}{a' Wa} \quad \text{w.r.t.} \quad a' Wa = 1
\]

(“a” is a unit vector)

The factors are obtained from the eigenvalues and eigenvector of \( W^{-1}B \).

The eigenvectors of \( W^{-1}B \) are the same as those of \( V^{-1}B \) → the factors are identical.

The eigenvalues are related with the following formula:

\[ \rho_m = \frac{\lambda_m}{1 - \lambda_m} \]

E.g. Bordeaux wine
With only the variables “temperature” and “sun”

<table>
<thead>
<tr>
<th>Root</th>
<th>Eigenvalue</th>
<th>Proportion</th>
<th>Canonical R</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.6432</td>
<td>0.9802</td>
<td>0.8518</td>
</tr>
<tr>
<td>2</td>
<td>0.0534</td>
<td>1</td>
<td>0.2251</td>
</tr>
</tbody>
</table>

E.g. The first factor explains 98% of the global between-class variation: 98% = \( \frac{2.6432}{2.6432 + 0.0534} \).

The two factors explain 100% of this variation \( [M = \min(2, 3-1) = 2] \)

⇒ The first factor is enough here!
Reading the results of DDA

Determining the right number of factors

Interpreting the factors
Descriptive Discriminant Analysis – Determining the right number of factors

We want to check

H0: the correlation ratios of the "q" last factors are zero
⇔ H0: $\eta_{K-q}^2 = \eta_{K-q-1}^2 = \cdots = \eta_{K-1}^2 = 0$
⇔ H0: we can ignore the “q” remaining factors

Test statistic

$$\Lambda_q = \prod_{m=K-q}^{K-1} \left(1 - \eta_m^2\right)$$

The lower is the value of LAMBDA, the more interesting are the factors.

In the case of Gaussian distribution (i.e. the data follows a multidimensional normal distribution in each group), we can use the Bartlett (chi-squared) or Rao transformation (Fisher).

<table>
<thead>
<tr>
<th>Root</th>
<th>Eigenvalue</th>
<th>Proportion</th>
<th>Canonical R</th>
<th>Wilks Lambda</th>
<th>CHI-2</th>
<th>d.f.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.6432</td>
<td>0.9802</td>
<td>0.8518</td>
<td>0.260568</td>
<td>41.0191</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.0534</td>
<td>1</td>
<td>0.2251</td>
<td>0.949308</td>
<td>1.5867</td>
<td>1</td>
<td>0.207802</td>
</tr>
</tbody>
</table>

The two first factors are together significant at 5% level; but the last factor is not significant alone.
**H0:** all the correlation ratio are zero

\[ \eta_1^2 = \cdots = \eta_{K-1}^2 = 0 \]

\( \Leftrightarrow \) **H0:** we cannot distinguish the groups centroid in the global representation space

\[
\Lambda = \prod_{m=1}^{K-1} \left(1 - \eta_m^2 \right)
\]

**MANOVA** test i.e. comparing multivariate means (centroids) of several groups

\[
H_0 : \begin{pmatrix} \mu_{1,1} \\ \vdots \\ \mu_{J,1} \\ \mu_{1,K} \\ \vdots \\ \mu_{J,K} \end{pmatrix} = \cdots = \begin{pmatrix} \mu_{1,1} \\ \vdots \\ \mu_{J,1} \\ \mu_{1,K} \\ \vdots \\ \mu_{J,K} \end{pmatrix}
\]

The lower is the value of LAMBDA, the more different are the centroids (0 ≤ \( \Lambda \) ≤ 1).

**Test statistic:**

*Wilks’ LAMBDA*

**Conclusion:** At least one centroid is different to the others.

**LAMBDA de Wilks = 0.26**

**Bartlett transformation**

\[ \text{CHI}^2 = 41.02 ; \text{p-value} < 0.0001 \]

**Rao transformation**

\[ F = 14.39 ; \text{p-value} < 0.0001 \]
Descriptive discriminant analysis – Interpreting the canonical variables (factors)
Standardized and unstandardized canonical coefficients

**Unstandardized coefficients**
These coefficients enables to calculate the canonical scores of the individuals (coordinates of the individuals, discriminant scores)

\[
Z = a_1(x_1 - \bar{x}_1) + \cdots + a_J(x_J - \bar{x}_J)
= a_0 + a_1x_1 + \cdots + a_Jx_J
\]

The unstandardized canonical coefficients do not allow to compare the influence of the variables because they are not defined on the same unit.

**Standardized coefficients**
These are the coefficients of the DDA on standardized variables. We can obtain the same values by multiplying the unstandardized coefficients with the pooled within-class standard deviation of the variables. The coefficients (influence) of the variables become comparable.

\[
\beta_j = a_j \times \sigma_j
\]

The pooled within class variance of the variable \(X_j\)

Standardized coefficients show the variable's contribution to calculating the discriminant score. Two correlated variables share their contribution, their true influence may be hidden (W.R. Klecka, “Discriminant Analysis”, 1980; page 33).

We must complete this analysis by studying the structure coefficients table.

Quality = DDA (Temperature, Sun) >>

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Coefficients</th>
<th>Unstandardized</th>
<th>Standardized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root n*1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Root n*2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temperature</td>
<td>0.007465</td>
<td>-0.009214</td>
<td>-0.653736</td>
</tr>
<tr>
<td>Sun</td>
<td>0.007479</td>
<td>0.010459</td>
<td>0.844002</td>
</tr>
<tr>
<td>constant</td>
<td>32.903185</td>
<td>16.049255</td>
<td></td>
</tr>
</tbody>
</table>

Ricco Rakotomalala
These are the bivariate correlation between the variables and the canonical variables. We can visualize the correlation circle such as for PCA (principal component analysis).

The 1st factor corresponds to the combination of high temperature and high periods of sunshine. The combination of high temperature and high periods of sunshine correspond to "good" wine.

These correlation coefficients allow to interpret easily the factors. If the sign are different to the standardized canonical coefficients \(\rightarrow\) collinearity between the variables.
Descriptive discriminant analysis – Interpreting the canonical variables (factors)
Within structure coefficients

These coefficients show how the variables are related to the canonical variable within the groups.

\[ r = 0.9334 \]

\[ r_w = 0.8134 \]

Often lower value than the total correlation (not always).

<table>
<thead>
<tr>
<th>Root Descriptors</th>
<th>Total</th>
<th>Within</th>
<th>Between</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>0.9334</td>
<td>0.8134</td>
<td>0.9949</td>
</tr>
<tr>
<td>Sun</td>
<td>0.9168</td>
<td>0.777</td>
<td>0.9934</td>
</tr>
</tbody>
</table>
Correlation of the variables with the factors by using only the group centroids.

Interesting but not always convenient. The value is +1 or -1 when we have only 2 groups (K = 2).

\[ r = 0.9334 \]

\[ r_B = 0.9949 \]

<table>
<thead>
<tr>
<th>Root Descriptors</th>
<th>Root n°1 Total</th>
<th>Root n°1 Within</th>
<th>Root n°1 Between</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>0.9334</td>
<td>0.8134</td>
<td>0.9949</td>
</tr>
<tr>
<td>Sun</td>
<td>0.9168</td>
<td>0.777</td>
<td>0.9934</td>
</tr>
</tbody>
</table>
Descriptive discriminant analysis – Interpreting the canonical variables (factors)

Group centroids into the discriminant representation space

Calculating the coordinates of the centroids in the new representation space.
This allows to identify the groups which are well highlighted.

The three groups are quite separate on the first factor
Nothing interesting on the second factor (low canonical correlation)

KIRSCH vs. the two other groups on the 1st factor
POIRE vs. MIRAB on the 2nd factor (significant canonical correlation)
Case study

Bordeaux wine (Tenenhaus, 2007; page 353)
Bordeaux wine - Description of the dataset

Some of the descriptors are correlated (see the correlation matrix).

(Red: Bad; blue: Medium; green: Good).
The groups are discernible, especially for some combination of variables.

The influence on the quality is not the same according to the variables.

There are outliers...

Correlation matrix

```R
> cor(wine[,1:4])

                      Temperature        Sun        Heat       Rain
Temperature  1.00000000  0.7123527  0.8650958 -0.4096188
Sun          0.71235274  1.0000000  0.6464478  0.4733991
Heat         0.86509584  0.6464478  1.0000000  0.4011372
Rain         -0.40961883 -0.4733991 -0.4011372  1.0000000
```
Bordeaux wine – Univariate analysis of the variables
Conditional distribution and correlation ratio

“Temperature”, “Sun” and “Heat” enable to well distinguish the groups. "Rain" seems less decisive.

For all the variables, the univariate one-way ANOVA (the class means are equal or not) is significant at 5% level.
Bordeaux wine – DDA results

Roots and Wilks’ Lambda

<table>
<thead>
<tr>
<th>Root</th>
<th>Eigenvalue</th>
<th>Proportion</th>
<th>Canonical R</th>
<th>Wilks Lambda</th>
<th>CHI-2</th>
<th>d.f.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.27886</td>
<td>0.95945</td>
<td>0.875382</td>
<td>0.205263</td>
<td>46.7122</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.13857</td>
<td>1</td>
<td>0.348867</td>
<td>0.878292</td>
<td>3.8284</td>
<td>3</td>
<td>0.280599</td>
</tr>
</tbody>
</table>

(a) The difference between groups is significant. (b) 96% of between-class variation is explained by the first factor. (c) The 2nd factor is not significant at 5% level, we can ignore it.

On the first factor, we observe the 3 groups. From the left to the right, we have the centroids of “good”, “medium” and “bad”.

The square of the correlation ratio for this factor is 0.766. This is higher than any univariate correlation ratio of the variables (the higher is "temperature" with $\eta^2 = 0.64$).
Bordeaux wine – Groups characteristics
Interpreting the canonical variables

Canonical Discriminant Function

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Unstandardized</th>
<th>Standardized</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Root n°1</td>
<td>Root n°2</td>
</tr>
<tr>
<td>Temperature</td>
<td>-0.008575</td>
<td>0.000046</td>
</tr>
<tr>
<td>Soleil</td>
<td>-0.006781</td>
<td>0.005335</td>
</tr>
<tr>
<td>Chaleur</td>
<td>0.027083</td>
<td>-0.127772</td>
</tr>
<tr>
<td>Pluie</td>
<td>0.005872</td>
<td>-0.006181</td>
</tr>
<tr>
<td>constant</td>
<td>32.911354</td>
<td>-2.167589</td>
</tr>
</tbody>
</table>

Factor Structure Matrix - Correlations

The first factor brings into opposition the “temperature” and the “sun” on the one side (high values: good wine), and the “rain” on the other side (high values: bad wine).

The influence of “heat” seems unclear. It has a positive influence on the first factor according to the canonical coefficients table. But it has a negative relation to the first factor according to the structure coefficients table.

Actually, this variable is highly correlated to “temperature”. The partial correlation ratio of “heat” by controlling “temperature” is very low (Tenenhaus, page 376) $\eta_{x,y/x_i}^2 = 0.0348$
Classifying an unseen instance

Using the results of DDA to determine the group membership of a new instance
Classification rule

Preamble

The linear (predictive) discriminant analysis (PDA) offers a more attractive theoretical framework for prediction, with explicit probabilistic assumptions.

Nevertheless, we can use the results of the DDA to classify individuals based on geometric rules.

Which group?

Steps:
1. As from the description of the individual, its coordinates in the discriminant dimensions are computed.
2. The distance to the conditional centroids is computed.
3. The instance is assigned to the group of which the centroid is the closest.
DDA from Temperature (X1) and Sun (X2) 

X1 = 3000 – X2 = 1100 – Year 1958 (based on the weather forecast)

1. Calculating the coordinates

\[ z_1 = 0.007457 \times x_1 + 0.007471 \times x_2 - 32.868122 \]
\[ = 0.007457 \times 3000 + 0.007471 \times 1100 - 32.868122 \]
\[ = -2.2780 \]

\[ z_2 = -0.009204 \times x_1 + 0.010448 \times x_2 + 16.032152 \]
\[ = -0.009204 \times 3000 + 0.010448 \times 1100 + 16.032152 \]
\[ = -0.0862 \]

2. Calculating the distance to the centroids

\[ d^2(bad) = (-2.2780 - (-1.8023))^2 + (-0.0832 - (-0.1538))^2 \]
\[ = 0.2309 \]
\[ d^2(good) = 18.1031 \]
\[ d^2(medium) = 5.3075 \]

3. Conclusion

The vintage 1958 has a high probability to be “bad”. It has a very low probability to be “good”.
Classifying a new instance
Euclidian distance into the discriminant dimensions = Mahalanobis distance into the initial representation space

We can obtain the same distance as preceding in the initial representation space by using the $W^{-1}$ metric: this is the Mahalanobis distance.

For the instance “1958”, we calculate its distance to the "bad" centroid as follows...

$$d^2(bad) = (x - \mu_{bad}) W^{-1} (x - \mu_{bad})$$

$$= (3000 - 3037.3 ; 1100 - 1126.4) \begin{pmatrix} 7668.46 & 1880.15 \\ 1880.15 & 6522.33 \end{pmatrix}^{-1} \begin{pmatrix} 3000 - 3037.3 \\ 1100 - 1126.4 \end{pmatrix}$$

$$= (-37.33 \ -26.42) \begin{pmatrix} 0.000140 & -0.000040 \\ -0.000040 & 0.000165 \end{pmatrix} \begin{pmatrix} -37.33 \\ -26.42 \end{pmatrix}$$

$$= 0.2309$$

$W = \begin{pmatrix} 7668.46 & 1880.15 \\ 1880.15 & 6522.33 \end{pmatrix}$ is the pooled within class SSCP matrix (sum of squares and cross products) [i.e. the covariance matrix multiplied by the degree of freedom (n-K)].

Why the results of DDA are important?

1. We have in addition an explanation of the prediction. "1958" is probably "bad" because of low temperature and low sun.

2. We can use only the significant canonical variables for the prediction. This is a kind of regularization (see "reduced rank LDA", Hastie et al., 2001).
Classifying an new instance
Specifying an explicit model

For an instance “i”, we calculate as follows its distance to the centroid of the group “k”. We take into account Q canonical variables (Q = M if we treat all the factors).

\[ d_i^2(k) = \sum_{m=1}^{Q} (z_{im} - \bar{z}_{m,k})^2 \]
\[ = \sum_{m=1}^{Q} z_{im}^2 + \bar{z}_{m,k}^2 - 2 z_{im} \bar{z}_{m,k} \]

\[ f_i(k) = \sum_{m=1}^{Q} \left( \bar{z}_{m,k} \times z_{im} - \frac{1}{2} \bar{z}_{m,k}^2 \right) \]
\[ = \sum_{m=1}^{Q} \bar{z}_{m,k} \times z_{im} - \frac{1}{2} \sum_{m=1}^{Q} \bar{z}_{m,k}^2 \]

Finding the closest centroid (minimization). We can transform it in a maximization problem by multiplying with -0.5

\[ k^* = \arg \min_k d_i^2(k) \iff k^* = \arg \max_k f_i(k) \]

Discriminant function for the factor “m”

\[ z_m = a_{0m} + a_{1m}x_1 + a_{2m}x_2 + \cdots + a_{jm}x_J \]

We have a linear classification function.

Finding the closest centroid (minimization). We can transform it in a maximization problem by multiplying with -0.5

\[ k^* = \arg \min_k d_i^2(k) \iff k^* = \arg \max_k f_i(k) \]

Discriminant function for the factor “m”

\[ z_m = a_{0m} + a_{1m}x_1 + a_{2m}x_2 + \cdots + a_{jm}x_J \]

E.g. Bordeaux wine with “temperature” (x_1)
and “sun” (x_2) – Only one factor (Q = 1)

For the instance (x_1 = 3000; x_2 = 1100)

\[ f(bad) = -1.8023 \times (0.007457x_1 + 0.007471x_2 - 32.868122) - \frac{1}{2}(-1.8023)^2 \]
\[ = -0.0134x_1 - 0.0135x_2 + 57.6129 \]
\[ f(good) = 0.0147x_1 + 0.0148x_2 - 66.9081 \]
\[ f(medium) = -0.0001x_1 - 0.0001x_2 + 0.3331 \]

Conclusion: the vintage “1958” will be probably « bad »

For the instance (x_1 = 3000; x_2 = 1100)

\[ f(bad) = 2.4815 \]
\[ f(good) = -6.5447 \]
\[ f(medium) = 0.0230 \]
Classification an new instance
What is the connection with the linear (predictive) discriminant analysis (PDA)?

The parametric linear discriminant analysis makes assumptions about the distribution and the dispersion of the observations (normal distribution, homogeneity of variances/covariances)

\[
d(Y_k, X) = \ln[P(Y = y_k)] + \mu_k \Sigma^{-1} X' - \frac{1}{2} \mu_k \Sigma^{-1} \mu_k'
\]

Classification rule from the DDA when we handle all the factors (M factors)

In conclusion, the classification rule of DDA is equivalent to the one of PDA if we have balanced class distribution i.e.

\[
P(Y = y_1) = \cdots = P(Y = y_K) = \frac{1}{K}
\]

Some tools make this assumption by default (e.g. default settings for the SAS PROC DISCRIM)

Introducing the correction derived from the estimated class distribution will improve the error rate (Hastie et al., 2001; page 95).
Some data mining tools

Tanagra, R and SAS
CANONICAL DISCRIMINANT ANALYSIS tool

The main results, usable for the interpretation, are available.

We can obtain the graphical representation of the individuals and the correlation circle for the variables (based on the total structure correlation).

French references use \( \frac{1}{n} \) for the estimation of the covariance.
DDA with TANAGRA
Graphical representation

Plotting the individuals into the discriminant dimensions

Correlation circle
The output is concise.

But with some programming instructions, we can obtain better. This is one of the main advantages of R.

English-speaking references use \( \frac{1}{(n-1)} \) for the estimation of the covariance.
DDA with R

With some programming instructions, the result is worth it …
DDA with SAS
The CANDISC procedure

Comprehensive results.

The “ALL” option allows to obtain all the intermediate results (matrices V, W, B; etc.).

English-speaking references use $[1/(n-1)]$ for the estimation of the covariance (such as $R$).
Conclusion
Conclusion

DDA: multivariate method for groups’ description and characterization

Tools for the interpretation of the results (test for significance of canonical variables, canonical coefficients, structure coefficients...)

Tools for the visualization of the results (individuals, variables)

The approach is related to other factorial methods (principal component analysis, canonical correlation)

The approach is in nature descriptive, but it can be implemented in a predictive framework easily.

The approach provides a white-box prediction (we can understand for which reason an unseen instance is assigned to such group).
References

