Discriminant Analysis for Interval Data

Paula Brito

Fac. Economia & LIAAD-INESC TEC, Universidade do Porto

ECI 2015 - Buenos Aires
T3: Symbolic Data Analysis:
Taking Variability in Data into Account

Outline

1. Discriminant analysis : Introduction

2. Discriminant analysis for Interval Data

Outline

1. Discriminant analysis : Introduction

2. Discriminant analysis for Interval Data

Linear Discriminant analysis : Objective

Applies to data described by numerical variables

The elements of the data set S are divided in groups/classes $C_1, \ldots, C_k$ **a priori**

- Identify the variables that best distinguish the groups ;
- Use these variables to form an indicator that allows representing, in a concise form, the differences between the groups ;
- Use the identified variables and the indicator to build a rule that allows classifying future cases in one of the two groups.

Linear Discriminant analysis : Solution

The discriminant function(s) is a linear combination of the descriptive variables

defined by the eigenvectors of $W^{-1}B$

W : matrix of sums of squares and cross-products within class

B : matrix of sums of squares and cross-products between classes

$$w_{jj'} = \sum_{\ell=1}^{k} \sum_{s_i \in C_\ell} (x_{ij} - \overline{x_j^{(\ell)}})(x_{ij'} - \overline{x_{j'}^{(\ell)}})$$

$$b_{jj'} = \sum_{\ell=1}^{k} n_\ell (\overline{x_j^{(\ell)}} - \overline{x_j})(\overline{x_{j'}^{(\ell)}} - \overline{x_{j'}})$$

Outline

1. Discriminant analysis : Introduction
2. Discriminant analysis for Interval Data

Discriminant analysis for Interval-valued variables

Three approaches (Duarte Silva, Brito (2006)):

- Assume an uniform distribution in each interval, and use measures proposed by Bertrand and Goupil (2000) and Billard and Diday (2003).
- Consider all the vertices of the hypercube representing each of the n individuals in the p-dimensional space, and perform a classical discriminant analysis of the resulting $n \times 2^p$ by p matrix.
- Represent each variable by the midpoints and ranges of its interval values, perform two separate classical discriminant analysis on these values and combine the results in some appropriate way, or analyze midpoints and ranges conjointly.

Discriminant analysis for Interval-valued variables

Uniformity-based approach

Problem with k groups :

$r = \min\{p, k-1\}$ new variables

collected in a $n \times r$ matrix $Z = Y \bigotimes \beta$

where β is the $p \times r$ matrix of the coefficients

Discriminant analysis for Interval-valued variables

Assuming uniformity in each observed interval

$m_j = \frac{1}{n} \sum_{i=1}^{n} \frac{l_{ij}+u_{ij}}{2} = \frac{1}{2}(\bar{l}_j + \bar{u}_j)$

is the mean value of the interval midpoints for variable Y_j.

The empirical variance is given by

$$s_j^2 = \int_{-\infty}^{+\infty} (\xi - m_j)^2 f_j(\xi) d\xi = \\ \frac{1}{3n} \sum_{i=1}^{n} (l_{ij}^2 + l_{ij} u_{ij} + u_{ij}^2) - \frac{1}{4n^2} \left[\sum_{i=1}^{n} (l_{ij} + u_{ij})\right]^2$$

and the empirical covariance :

$$s_{jj'} = cov(Y_j, Y_{j'}) = \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (\xi_1 - m_j)(\xi_2 - m_{j'}) f_{jj'}(\xi_1, \xi_2) d\xi_1 d\xi_2 = \\ \frac{1}{4n} \sum_{i=1}^{n} [(l_{ij} + u_{ij})(l_{ij'} + u_{ij'})] - \frac{1}{4n^2} \left[\sum_{i=1}^{n} (l_{ij} + u_{ij})\right] \left[\sum_{i=1}^{n} (l_{ij'} + u_{ij'})\right]$$

Discriminant analysis for Interval-valued variables

If the n observations are partitioned into k groups, $C_1, \ldots, C_k$

The global empirical density functions are mixtures of the corresponding group specific functions

It is proved that s_j^2 and $s_{jj'}$ are decomposed in a within group component and a between group component.

The terms of the matrices W and B are given by:

$$w_{jj} = \frac{1}{3}\sum_{i=1}^{n}(u_{ij}^2 + u_{ij}l_{ij} + l_{ij}^2) - \sum_{\alpha=1}^{k} n_\alpha m_{\alpha j}^2$$

$$w_{jj'} = \frac{1}{4}\sum_{i=1}^{n}(u_{ij} + l_{ij})(u_{ij'} + l_{ij'}) - \sum_{\alpha=1}^{k} n_\alpha m_{\alpha j} m_{\alpha j'},\ j \neq j^{\,'}$$

$$b_{jj'} = \sum_{\alpha=1}^{k} n_\alpha m_{\alpha j} m_{\alpha j'} - n m_j m_{j'}$$

for $j, j^{'} = 1, \ldots, p$.

Discriminant analysis for Interval-valued variables

As in the classical case, the discriminant functions coefficients are given by the eigenvectors of the product $W^{-1}B$.

Single point representations on a discriminant space are given directly and interval representations may be determined.

Discriminant analysis for Interval-valued variables

Vertices method

Interval matrix $I \rightarrow$ new matrix of single real values M

each row i of I gives rise to 2^p rows of M

corresponding to all possible combinations of the limits of intervals $[l_{ij}, u_{ij}], j = 1, \ldots, p$.

Classical discriminant analysis on matrix $M \rightarrow$

factorial representation of points, one for each of the 2^p vertices.

Discriminant analysis for Interval-valued variables

A representation in the form of intervals may be obtained, as for PCA:

- Let Q_i be the set of row indices q in matrix M which refer to the vertices of the hypercube corresponding to s_i.
- For $q \in Q_i$ let $\zeta_{q\ell}$ be the value of the ℓ-th real-valued discriminant function for the vertex with row index q.
- The value of the ℓ-th interval discriminant variate z for s_i, is then defined by

$$\begin{aligned} \underline{z}_{i\ell} &= \text{ Min } \{\zeta_{q\ell}, q \in Q_i\} \\ \overline{z}_{i\ell} &= \text{ Max } \{\zeta_{q\ell}, q \in Q_i\} \end{aligned}$$

These variates may be used for description purposes as well as for classification

Discriminant analysis for Interval-valued variables

Centres and ranges approach

- Represent observed intervals by midpoints and ranges
- Perform two separate classical discriminant analysis on these values
- Combine the results in some appropriate way

Alternatively:

Combined discriminant analysis performed simultaneously for midpoints and ranges.

This is particularly relevant when midpoints and ranges are related in such a way that their contribution to group separation cannot be recovered by two independent analysis.

Discriminant analysis for Interval-valued variables

Allocation rules are based on point distances or distances between intervals, accordingly.

Distributional approach:

Allocating each observation to the group with nearest centroïd in the discriminant space, according to a simple Euclidean distance.

Distinct prior probabilities and/or misclassification costs may be taken into account as in the classical case.

Linear combinations of the interval variables may be determined, that produce interval-valued discriminant variates.

In this case, allocation rules may be derived by using distances between interval vectors, e.g., Hausdorff distance.

Discriminant analysis for Interval-valued variables

Vertices approach: Discriminant variates are interval-valued, so this same allocation rule is applied.

Midpoints and ranges approach:

only point distances are used to define allocation rules.

Two different situations occur:

- two separate analysis for midpoints and ranges → generally the discriminant variates are correlated → Mahalanobis distance
- single discriminant analysis taking into account both midpoints and ranges → Euclidean distance

Application

"Car" data set

33 car models described by 8 interval variables:
Price, Engine Capacity, Top Speed, Acceleration, Step, Length, Width, Height, considered as descriptive variables

A nominal variable *Category* is used as a *a priori* classification.

	Price	Engine Capacity	. . .	Height	Category
Alfa 145	$[27806, 33596]$	$[1370, 1910]$	. . .	$[143, 143]$	Utilitarian
Alfa 156	$[41593, 62291]$	$[1598, 2492]$	. . .	$[142, 142]$	Berlina
. . .	. . .	. . .	. . .	. . .	. . .
Porsche 25	$[147704, 246412]$	$[3387, 3600]$	. . .	$[130, 131]$	Sportive
Rover 25	$[21492, 33042]$	$[1119, 1994]$	. . .	$[142, 142]$	Utilitarian
Passat	$[39676, 63455]$	$[1595, 2496]$	. . .	$[146, 146]$	Luxury

Application

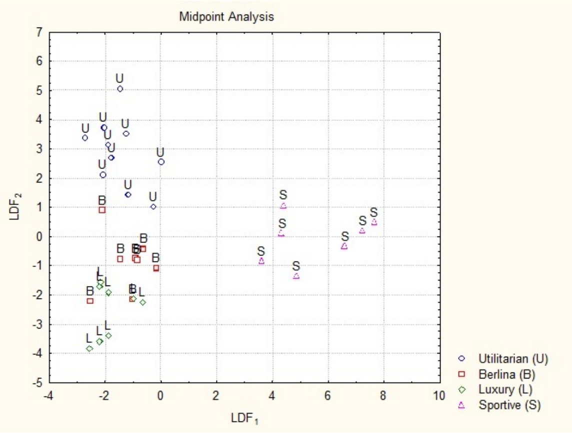

Figure: Point representations - distributional approach

Application

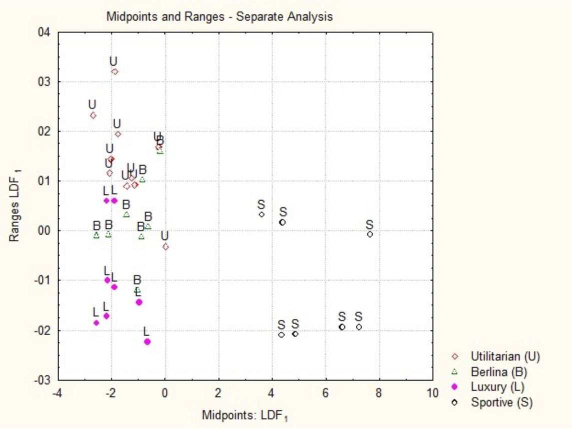

Figure: Point representations - midpoints and ranges separate analysis

Application

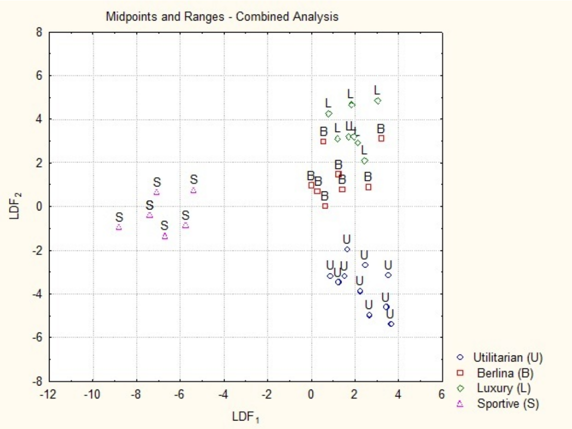

Figure: Point representations - midpoints and ranges combined analysis

Discriminant analysis: other approaches

SFDA: Symbolic Factorial Discriminant Analysis
(Lauro, Verde, Irpino (2008))

- SODAS module
- Performed on a set of entities, belonging to a priori r classes, identified by a nominal descriptor
- Entities can be described by any kind symbolic variable: quantitative single, interval, categorical single, categorical multi-valued and modal
- Mixed variables, logical dependence rules, taxonomies and missing values (NULL) are admitted
- The classification rule in SFDA is a geometrical one and it is performed according a proximity measure

Discriminant analysis: other approaches

The method consists of the following steps:

- Quantification of the descriptors
- FDA on the quantified predictors
- Symbolic interpretation of the results (factorial discriminant axes and classification rules) according to the nature data

Discriminant analysis: other approaches

First phase: numerical coding in a table Z_{ij}.

The chosen system of coding differs according to the type of descriptor.

In particular, if Y_j is:

- categorical multi-valued variable - the coding table is a binary matrix of values (0/1); s_i is coded with k_{ij} rows of Z_ij, being k_{ij} the categories of Y_j for s_i
- modal variable - s_i is coded in the table Z_{ij} according to the frequencies or probabilities that it assumes for the categories of Y_j
- a “quantitative single” or “interval” variable - it is categorized and codified according to a fuzzy system of coding (e.g. using Basic splines functions)

C_{nr} : indicator matrix with respect to the *r a prori* classes

The factorial discriminant axes are obtained as solutions of

$$\left(\tilde{\Phi}^t H \tilde{\Phi}\right)^{-1} \left(\tilde{\Phi}^t H C\right) \left(C^t H C\right)^{-1} \left(C^t H \tilde{\Phi}\right) \nu_m = \lambda_m \nu_m$$