Linear Models for Complex Data Analysis

Paula Brito

FEP & LIAAD - INESC TEC, Universidade do Porto, Portugal

Joint work with Sónia Dias and Paula Amaral

Optimization, Game Theory, and Data Analysis
December 20-21, 2018, University of Vienna, Austria

The starting point

8^{th} Workshop of the Working Group
"Matrix Computations and Statistics"
a satellite meeting of COMPSTAT 2006
Salerno, Italy, 2-3 September 2006

Outline

Outline

The data

Classical data analysis :
Data is represented in a $n \times p$ matrix
each of n individuals (in row) takes one single value
for each of p variables (in column)

	Nb. passengers	Delay (min)	Airline	Aircraft
Flight 1	200	20	Air France	Airbus
Flight 2	120	0	Ryanair	Boeing
Flight 3	100	10	Lufthansa	Airbus

The data

Symbolic Data Analysis :
to take into account **variability** inherent to the data

Variability occurs when we have

- Descriptors on flights, but: analyse the airline companies - not each individual flight
- Descriptors on prescriptions, but: analyse patients, or doctors - not the individual prescriptions
- Official statistics - Descriptors on citizens, but: analyse the cities, the regions - not the individual citizens

$\Longrightarrow$ (symbolic) variable values are

sets, intervals

distributions on an underlying set of sub-intervals or categories

Micro-data $\longrightarrow$ **Macro-data**

The data

Example : Data for three airline companies (e.g. arrival flights)

Airline	Nb Passengers	Delay (min)	Aircraft
A	180	10	Boeing
B	120	0	Boeing
A	200	20	Airbus
C	80	15	Embraer
B	100	5	Embraer
A	300	35	Airbus
C	70	30	Embraer
...	...	...	...

↓

Airline	Nb. Passengers	Delay (min)	Aircraft
A	[180, 300]	{[0, 10[, 0.33; [10, 30[, 0.33; [30, 60], 0.33}	{Airbus (2/3), Boeing (1/3)}
B	[100, 120]	{[0, 10[, 1.0; [10, 30[, 0; [30, 60], 0}	{Boeing (1/2), Embraer (1/2)}
C	[70, 80]	{[0, 10[, 0; [10, 30[, 0.5; [30, 60[, 0.45; [60, 90], 0.05}	{Embraer (1)}

The data

- In most common applications, symbolic data arises from the aggregation of micro data
- Often reported as such: temperature min-max intervals , financial assets daily min-max or open-close values
- They also occur directly, in descriptions of concepts : diseases, biological species (plants, etc.), technical specifications,...
- Quantile lists: infant growth, plant measures, etc.

Brito, P. (2014). Symbolic Data Analysis: Another Look at the Interaction of Data Mining and Statistics. *WIREs Data Mining and Knowledge Discovery*, 4(4), 281–295.

Symbolic Variable types

- Numerical (Quantitative) variables
  - Numerical single-valued variables
  - Numerical multi-valued variables
  - **Interval variables**
  - **Distributional variables: Histograms, Quantile lists**
- Categorical (Qualitative) variables :
  - Categorical single-valued variables
  - Categorical multi-valued variables
  - Distributional variables : Categorical modal - Compositions

The data

Airline	Nb. Passengers	Delay (min)	Aircraft
A	[180, 300]	{[0, 10[, 0.33; [10, 30[, 0.33; [30, 60], 0.33}	{Airbus (2/3), Boeing (1/3)}
B	[100, 120]	{[0, 10[, 1.0; [10, 30[, 0; [30, 60], 0}	{Boeing (1/2), Embraer (1/2)}
C	[70, 80]	{[0, 10[, 0; [10, 30[, 0.5; [30, 60[, 0.45; [60, 90], 0.05}	{Embraer (1)}

Outline

1. Variability in Data
2. Histogram-valued variables
3. Linear Regression for histogram data
4. Discriminant Analysis with histogram data
5. Summary and References

Histogram-valued variables

Histogram-valued variable : $Y : S \rightarrow B$

B : set of probability or frequency distributions over a set of sub-intervals

$$Y(s_i) = (I_{i1}, p_{i1}; \ldots ; I_{ik_i}, p_{iK_i})$$

$p_{i\ell}$: probability or frequency associated to $I_{i\ell} = [\underline{I}_{i\ell}, \overline{I}_{i\ell}[$
$p_{i1} + \ldots + p_{iK_i} = 1$

$Y(s_i)$ may be represented by the histogram :

$$H_{Y(s_i)} = ([\underline{I}_{i1}, \overline{I}_{i1}[, p_{i1}; \ldots ; [\underline{I}_{iK_i}, \overline{I}_{iK_i}], p_{ijK_i})$$

Histogram data

	Y_1	...	Y_p
s_1	$\{[\underline{I}_{111}, \overline{I}_{111}[, p_{111}; \ldots; [\underline{I}_{11K_{11}}, \overline{I}_{11K_{11}}], p_{11K_{11}}\}$	...	$\{[\underline{I}_{1p1}, \overline{I}_{1p1}[, p_{1p1}; \ldots; [\underline{I}_{1pK_{1p}}, \overline{I}_{1pK_{1p}}], p_{1pK_{1p}}\}$
...	...		...
s_i	$\{[\underline{I}_{i11}, \overline{I}_{i11}[, p_{i11}; \ldots; [\underline{I}_{i1K_{i1}}, \overline{I}_{i1K_{i1}}], p_{i1K_{i1}}\}$	...	$\{[\underline{I}_{ip1}, \overline{I}_{ip1}[, p_{ip1}; \ldots; [\underline{I}_{ipK_{ip}}, \overline{I}_{ipK_{ip}}], p_{ipK_{ip}}\}$
...	...		...
s_n	$\{[\underline{I}_{n11}, \overline{I}_{n11}[, p_{n11}; \ldots; [\underline{I}_{n1K_{n1}}, \overline{I}_{n1K_{n1}}], p_{n1K_{n1}}\}$	...	$\{[\underline{I}_{np1}, \overline{I}_{np1}[, p_{np1}; \ldots; [\underline{I}_{npK_{np}}, \overline{I}_{npK_{np}}], p_{npK_{np}}\}$

Histogram-valued variables

- Assumption : within each sub-interval $[\underline{I}_{ij\ell}, \overline{I}_{i\ell}[$ the values of variable Y_j for observation s_i, are uniformly distributed
- For each variable Y_j the number and length of sub-intervals in $Y_j(s_i)$, $i = 1, \ldots, n$ may be different
- Interval-valued variables : particular case of histogram-valued variables: $Y_j(s_i) = [l_{ij}, u_{ij}] \rightarrow H_{Y_j(s_i)} = ([l_{ij}, u_{ij}], 1)$

Histogram-valued variables

$Y(s_i)$ can, alternatively, be represented by the inverse cumulative distribution function - quantile function

$\Psi^{-1} : [0,1] \longrightarrow \mathbb{R}$

$$\Psi_i^{-1}(t) = \begin{cases} \underline{I}_{i1} + \frac{t}{w_{i1}} r_{i1} \ \ if \ \ 0 \leq t < w_{i1} \\ \underline{I}_{i2} + \frac{t - w_{i1}}{w_{i2} - w_{i1}} r_{i2} \ \ if \ \ w_{i1} \leq t < w_{i2} \\ \vdots \\ \underline{I}_{ijK_i} + \frac{t - w_{iK_i - 1}}{1 - w_{iK_i - 1}} r_{iK_i} \ \ if \ \ w_{iK_i - 1} \leq t \leq 1 \end{cases}$$

where $\quad w_{ih} = \sum_{\ell=1}^{h} p_{i\ell}, h = 1, \ldots, K_i; r_{i\ell} = \overline{I}_{i\ell} - \underline{I}_{i\ell}$
for $\ell = \{1, \ldots, K_i\}$.

These are piecewise linear functions.

Histogram-valued variables: Example

Studying the performance of some administrative offices - time people have to wait before being taken care of:

Office	Waiting Times (minutes)
A	5, 10, 15, 17, 20, 20, 25, 30, 30, 32, 35, 40, 40, 45, 50, 50
B	5, 8, 10, 12, 15, 20, 25, 25, 30, 32, 35, 35, 45, 52, 55, 60

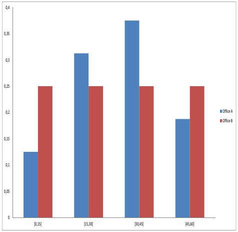
Average waiting time : 29.0 minutes for both offices

Description in terms of histograms :

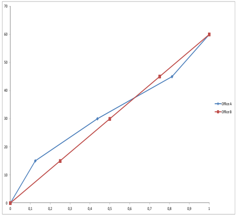
Office	Waiting Times (minutes)
A	$\{[0, 15[, 0.125; [15, 30[, 0.3125; [30, 45[, 0.375; [45, 60], 0.1875\}$
B	$\{[0, 15[, 0.25; [15, 30[, 0.25; [30, 45[, 0.25; [45, 60], 0.25\}$

Histogram-valued variables: Example

Histograms :



Quantile functions :



$$\Psi^{-1}(t) = \begin{cases} 120t & \text{if } 0 \leq t \leq 0.125 \\ 48t + 9 & \text{if } 0.125 \leq t \leq 0.4375 \\ 40t + 12.5 & \text{if } 0.4375 \leq t \leq 0.8125 \\ 80t - 20 & \text{if } 0.8125 \leq t \leq 1 \end{cases}$$

$\Psi^{-1}(t) = 60t$ for $0 \leq t \leq 1$

Histogram-valued variables: Distance measures

Many measures proposed in the literature
(see e.g. Bock and Diday (2000), Gibbs (2002))

Divergency measures	
Kullback-Leibler	$D_{KL}(f,g) = \int_{\mathbb{R}} log\left(\frac{f(x)}{g(x)}\right) f(x)dx$
Jeffreys	$D_J(f,g) = D_{KL}(f,g) + D_{KL}(g,f)$
χ^2	$D_{\chi^2}(f,g) = \int_{\mathbb{R}} \frac{\lvert f(x)-g(x)\rvert^2}{g(x)} dx$
Hellinger	$D_H(f,g) = \left[\int_{\mathbb{R}} \left(\sqrt{f(x)} - \sqrt{g(x)}\right) dx\right]^{\frac{1}{2}}$
Total variation	$D_{var}(f,g) = \int_{\mathbb{R}} \lvert f(x)-g(x)\rvert dx$
Kolmogorov	$D_K(f,g) = \max_{\mathbb{R}} \lvert F(x)-G(x)\rvert$
Wasserstein	$D_W(f,g) = \int_0^1 \lvert F^{-1}(t) - G^{-1}(t)\rvert dt$
Mallows	$D_M(f,g) = \sqrt{\int_0^1 (F^{-1}(t) - G^{-1}(t))^2\, dt}$

Histogram-valued variables: Distance measures

- **Wasserstein distance** :
  $D_W(\Psi^{-1}_{Y(i)}, \Psi^{-1}_{Y(i')}) = \int_0^1 \left| \Psi^{-1}_{Y(i)}(t) - \Psi^{-1}_{Y(i')}(t) \right| dt$

- **Mallows distance**:
  $D_M(\Psi^{-1}_{Y(i)}, \Psi^{-1}_{Y(i')}) = \sqrt{\int_0^1 (\Psi^{-1}_{Y(i)}(t) - \Psi^{-1}_{Y(i')}(t))^2 dt}$

Under the uniformity hypothesis,
and considering a fixed weight decomposition
(same weights, different intervals),
we have (Irpino and Verde, 2006):

$$D^2_M(\Psi^{-1}_{Y(i)}, \Psi^{-1}_{Y(i')}) =$$
$$= \sum_{\ell=1}^{K} p_\ell \left[(c_{Y(i)} - c_{Y(i')})^2 + \frac{1}{3}(r_{Y(i)} - r_{Y(i')})^2 \right]$$

Histogram-valued variables: Distance measures

- **Squared Euclidean distance**

$$d_E^2(Y_i, Y_{i'}) = \sum_{\ell=1}^{k} (p_{i\ell} - p_{i'\ell})^2$$

Differences between weights, fixed partition
(same intervals for all observations)

Descriptive Statistics for Histogram Variables

Irpino and Verde (2015):
Basic statistics obtained using a metric-based approach

Fréchet Mean :

$$M = \underset{x}{argmin} \sum_{i=1}^{n} w_i d^2(s_i, x)$$

Barycenter

Euclidean distance : mean distribution or barycenter is the finite uniform mixture of the given distributions

Descriptive Statistics for Histogram Variables: Barycenter

Mallows distance : mean distribution or barycenter obtained from the mean quantile function

The Mallows **barycentric histogram** is the solution of the minimization problem

$$\min \quad \sum_{i=1}^{n} D_M^2(\Psi_{Y(i)}^{-1}(t), \Psi_{Y_b}^{-1}(t))$$

that is, the quantile function where the centers and half ranges of each subinterval ℓ are the classical mean of the centers and half ranges of all observations

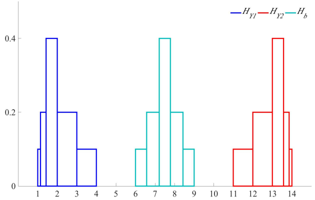
Need to re-write the histograms - and quantile functions - with the same weight decomposition

Descriptive Statistics for Histogram Variables: Barycenter

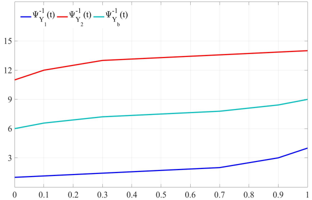
$H_1 = \{[1; 2[; 0.7; [2; 3[; 0.2; [3; 4]; 0.1\}$
$H_2 = \{[11; 12[; 0.1; [12; 13[; 0.2; [13; 14]; 0.7\}$

Barycentric histogram:
$H_b = \{[6; 6.58[; 0.1; [6.58; 7.21[; 0.2;$
$[7.21; 7.79[; 0.4; [7.79; 8.43[; 0.2[8.43; 9]; 0.1\}$

Histograms :



Quantile functions :



Histogram-valued variables: Mallows distance properties

Given a partition in k groups, the Mallows distance fulfils the Huygens theorem decomposition in Between and Within dispersion (Irpino and Verde, 2006):

$$\sum_{i=1}^{n} D_M^2(\Psi_{s_i}^{-1}(t), \overline{\Psi_S^{-1}}(t)) =$$
$$\sum_{h=1}^{k} n_h D_M^2(\overline{\Psi_S^{-1}}(t), \overline{\Psi_{C_h}^{-1}}(t)) +$$
$$+ \sum_{h=1}^{k} \sum_{i \in C_h} D_M^2(\Psi_{s_i}^{-1}(t), \overline{\Psi_{C_h}^{-1}}(t))$$

where n_h is the number of observations in group C_h

Irpino A., Verde R. (2006). A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. *Data Science and Classification, Proc. IFCS 2006.* Springer, 185-192

Outline

First linear regression models

- First linear regression method for histogram-valued data due to Billard and Diday (2006)
  - Model based on the - real-valued - first and second-order moments for histogram-valued variables obtained previously
  - From these, the regression coefficients are derived
- Irpino and Verde (2008) developed a linear regression model
  - Minimizing the Mallows's distance between the observed and the derived quantile functions of the dependent variable
  - The method relies on the exploitation of the properties of a decomposition of the Mallows's distance
  - Used to measure the sum of squared errors and rewrite the model
  - Splitting the contribution of the predictors in a part depending on the averages of the distributions and another depending on the centered quantile distributions

Distribution and Symmetric Distribution Linear Regression model

Joint work with Sónia Dias (IPVC & INESC TEC)

- Dias and Brito (2015) propose a new Linear Regression model for histogram-valued variables
- Distributions are represented by their quantile functions
- The model includes both the quantile functions that represent the distributions that the independent histogram-valued variables take, and the quantile functions that represent the distributions that the respective symmetric histogram-valued variables take - two terms per independent variable

Linear combination of quantile functions

The linear combination of quantile functions **is not defined** as:

$$\Psi^{-1}_{Y(i)}(t) = a_1\Psi^{-1}_{X_1(i)}(t) + a_2\Psi^{-1}_{X_2(i)}(t) + \ldots + a_p\Psi^{-1}_{X_p(i)}(t)$$

- Because when we multiply a quantile function by a negative number we do not obtain a non-decreasing function
- If non-negativity constraints are imposed on the parameters a_j, $j \in \{1, 2, \ldots, p\}$ a quantile function is always obtained.
  However, this solution forces a direct linear relation between $\Psi^{-1}_{Y(i)}(t)$ and $\Psi^{-1}_{X_j(i)}(t)$
- Dias and Brito (2015) proposed a definition for linear combination of quantile functions that solves the problem of the semi-linearity of the space of the quantile functions

Dias, S. and Brito, P. (2015), Linear Regression Model with Histogram-Valued Variables. *Statistical Analysis and Data Mining*, 8(2),75-113

Definition of linear combination

To allow for a direct and an inverse linear relation between the quantile functions, the linear combination includes:

- $\Psi^{-1}_{X_j}(t)$ that represents the distributions of the histogram-valued variables X_j
- $-\Psi^{-1}_{X_j}(1-t)$ the quantile function that represents the respective symmetric histograms.

Linear combination between quantile functions

The quantile function Ψ^{-1}_{Y} may be expressed as a linear combination of $\Psi^{-1}_{X_j}(t)$ and $-\Psi^{-1}_{X_j}(1-t)$ as follows:

$$\Psi^{-1}_{Y}(t) = \sum_{j=1}^{p} a_j \Psi^{-1}_{X_j}(t) - \sum_{j=1}^{p} b_j \Psi^{-1}_{X_j}(1-t) + \gamma$$

with $t \in [0,1]$; $a_j, b_j \geq 0$, $j \in \{1,2,\ldots,p\}$.

Distribution and Symmetric Distribution Linear Regression model

- Non-negativity restrictions on the parameters do not imply a direct linear relationship
- Uses the Mallows distance to quantify the error
- Determination of the model requires solving a quadratic optimization problem, subject to non-negativity constraints on the unknowns

Distribution and Symmetric Distribution Linear Regression model

The parameters of the model are an optimal solution of the minimization problem:

Minimize $$SE = \sum_{i=1}^{n} D_M^2(\Psi^{-1}_{Y(i)}, \Psi^{-1}_{\widehat{Y}(i)})$$

with $a_j, b_j \geq 0$, $j = \{1, 2, \ldots, p\}$ and $\gamma \in \mathbb{R}$

$\longrightarrow$ Kuhn Tucker optimality conditions allow defining a measure to evaluate the quality of fit of the model (determination coefficient), Ω

Distribution and Symmetric Distribution Linear Regression model

- Experiments, both with small real data sets and simulated data: the model works well
- The goodness-of-fit measure shows good behaviour

Alternative version of the model has been developed:

- The constant term is itself a distribution (not a real number)
- Allows for a better interpretation of the obtained model coefficients

Models studied for the special case of interval-valued variables, with extension to triangular distributions within intervals:

Dias, S. and Brito, P. (2017). Off the Beaten Track: A New Linear Model for Interval Data. *European Journal of Operational Research*, 258(3), 1118–1130.

Distributional Data : Crimes in USA regression model

Original data: Socio-economic data from the '90 Census Crime data from 1995

First level units: Cities of the USA states

Original variables:

- Response variable: Y = (Log) total number of violent crimes per 100 000 habitants (LVC)
- Four explicative variables:
  - X1 = percentage of people aged 25 and over with less than 9th grade education
  - X2 = percentage of people aged 16 and over who are employed
  - X3 = percentage of population who are divorced
  - X4 = percentage of immigrants who immigrated within the last 10 years

Distributional Data : Crimes in USA regression model

Contemporary aggregation per state $\rightarrow$
Higher level units: USA states; 20 states considered

Observations associated to each unit:
The distributions of the records of the cities of the respective state

Response histogram-valued variable LVC :
distributions of the log of the number of violent crimes for each state

Distributional Data : Crimes in USA regression model

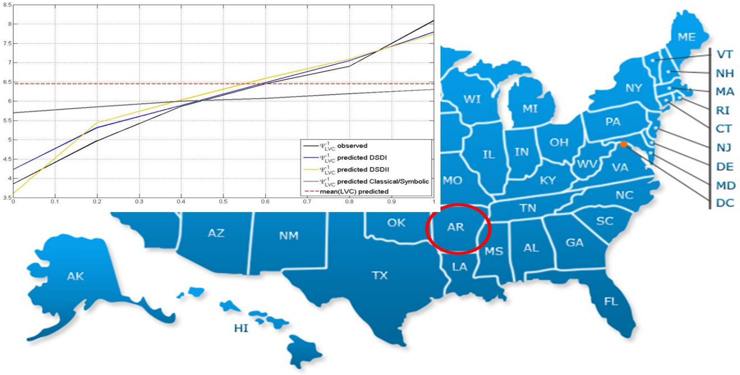
Model DSD I:

$$\Psi^{-1}_{\widehat{LVC}(j)}(t) = 3.9321 + 0.0009\Psi^{-1}_{X_1(j)}(t) - 0.0123\Psi^{-1}_{X_2(j)}(1-t) +$$
$$+0.2073\Psi^{-1}_{X_3(j)}(t) - 0.0353\Psi^{-1}_{X_3(j)}(1-t) + 0.0187\Psi^{-1}_{X_4(j)}(t); t \in [0,1]$$

Goodness-of-fit measure : $\Omega = 0.87$

X1, X3 and X4 : direct influence in the logarithm of the number of violent crimes
X2 (percentage of employed people) : opposite effect

Distributional Data : Crimes in USA regression model



$$H_{LVC}(AR) = \{[4.2250, 5.3158), 0.2; [5.3158, 5.8887), 0.2; [5.8887, 6.4802), 0.2; [6.4802, 7.0509), 0.2; [7.0509, 7.7913], 0.2\}$$

Outline

Linear Discriminant Analysis

Joint work with Sónia Dias (IPVC & INESC TEC) & Paula Amaral (NOVA Univ. of Lisbon)

Let S be partitioned in k groups, $G_h, h = 1, \ldots, k$.

A linear discriminant function is a linear combination of the explicative variables :

$$\Psi^{-1}_{D(i)}(t) = \sum_{j=1}^{p} a_j(\Psi^{-1}_{X_j(i)}(t) - \overline{\Psi^{-1}_{X_j}}(t)) +$$

$$+ \sum_{j=1}^{p} b_j(-\Psi^{-1}_{X_j(i)}(1-t) + \overline{\Psi^{-1}_{X_j}}(1-t)) \quad \text{with } a_j,\ b_j \geq 0$$

Alternatively: $\Psi^{-1}_{D(i)}(t) = \Psi^{-1}_{S(i)}(t) - \overline{\Psi^{-1}_{S}}(t)$ where

$$\Psi^{-1}_{S(i)}(t) = \sum_{j=1}^{p} a_j \Psi^{-1}_{X_j(i)}(t) - b_j \Psi^{-1}_{X_j(i)}(1-t)$$

$$\overline{\Psi^{-1}_{S}}(t) = \sum_{j=1}^{p} a_j \overline{\Psi^{-1}_{X_j}}(t) - b_j \overline{\Psi^{-1}_{X_j}}(1-t) \quad , \ a_j, b_j \geq 0$$

Discriminant Function

Classical Model

Discriminant Function

$$S(i) = \sum_{j=1}^{p} \gamma_j x_j(i)$$

The weight vector γ is obtained such that:

- the ratio of the **variability between groups** relatively to the **variability within groups** is maximized

$$\lambda = \frac{\gamma' \mathbf{B} \gamma}{\gamma' \mathbf{W} \gamma}$$

where
B - matrix of the sum of the squares between-groups
W - matrix of the sum of the squares within-groups

Symbolic Model

Discriminant Function

$$\Psi^{-1}_{S(i)}(t) = \sum_{j=1}^{p} a_j \Psi^{-1}_{X_j(i)}(t) - \sum_{j=1}^{p} b_j \Psi^{-1}_{X_j(i)}(1-t)$$

with $a_j, \ b_j \geq 0$.

The weight vector $\gamma \geq 0$ is obtained such that:

- the ratio of the **variability between groups** relatively to the **variability within groups** is maximized

$$\lambda = \frac{\gamma' \mathbf{B} \gamma}{\gamma' \mathbf{W} \gamma}$$

- The evaluation of the variability between scores is based on the Mallows distance

Discriminant Function

Classical Model

Decomposition of the matrix of the Sums of Squares and Cross-Products (SSCP):

$$\mathbf{T} = \mathbf{B} + \mathbf{W}$$

B - matrix of the sum of squares and cross-products between-groups
W - matrix of the sum of squares and cross-products within-groups

Consequently:

$$\gamma' \mathbf{T} \gamma = \gamma' (\mathbf{B} + \mathbf{W}) \gamma = \gamma' \mathbf{B} \gamma + \gamma' \mathbf{W} \gamma$$

$$\gamma' \mathbf{T} \gamma = \sum_{i=1}^{n} d^2(S(i), \overline{S})$$

with $S(i) = \sum_{j=1}^{p} \gamma_j x_j(i)$ and $\overline{S} = \frac{1}{n} \sum_{i=1}^{n} S(i)$

Symbolic Model

Sum of the squares of the Mallows distance between $\Psi^{-1}_{S(i)}(t)$ and $\overline{\Psi^{-1}_{S}}(t)$,

$$\sum_{i=1}^{n} D^2_M(\Psi^{-1}_{S(i)}(t), \overline{\Psi^{-1}_{S}}(t)) = \gamma' \mathbf{T} \gamma$$

According to the Huygens theorem :

$$\sum_{i=1}^{n} D^2_M(\Psi^{-1}_{S(i)}(t), \overline{\Psi^{-1}_{S}}(t)) = \sum_{h=1}^{k} |G_h| D^2_M(\overline{\Psi^{-1}_{S}}(t), \overline{\Psi^{-1}_{S_h}}(t)) + + \sum_{h=1}^{k} \sum_{i \in G_h} D^2_M(\Psi^{-1}_{S(i)}(t), \overline{\Psi^{-1}_{S_h}}(t))$$

with $\overline{\Psi^{-1}_{S_h}}(t) = \sum_{j=1}^{p} \left[a_j \overline{\Psi^{-1}_{X_{jh}}}(t) - b_j \overline{\Psi^{-1}_{X_{jh}}}(1-t) \right]$

In matricial notation:

$$\gamma' \mathbf{T} \gamma = \gamma' \mathbf{B} \gamma + \gamma' \mathbf{W} \gamma$$

T, **B**, **W** are $m \times m$ matrices, $m = 2p$.

Discriminant Function

Classical Model

Optimization problem:

Maximize the ratio

$$\lambda = \frac{\gamma' \mathbf{B} \gamma}{\gamma' \mathbf{W} \gamma}$$

Goal: Estimate vector γ such that the variability of the scores is maximal between groups and minimal within groups.

Complexity of the optimization problem:
- Easy to find the optimal solution

Symbolic Model

Optimization problem:

Maximize the ratio

$$\lambda = \frac{\gamma' \mathbf{B} \gamma}{\gamma' \mathbf{W} \gamma}$$

subject to $\gamma \geq 0$

Optimization of rational quadratic functions

- Hard optimization problem
- Nonconvex
- Easy to find a good solution
- Difficult to prove optimality
- Global optimal certificate of solution provided by BARON was only possible using a copositive relaxation

Conic Optimization

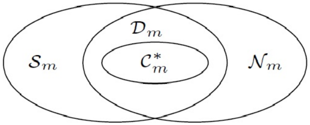
Optimization problem of the discriminant method:

$$\varphi = max \left\{ f(x) = \frac{x'\mathbf{B}x}{x'\mathbf{W}x} \ : x \in \mathbb{R}^m_+ \right\} = max \left\{ \mathbf{B} \cdot X : \ \mathbf{W} \cdot X = 1, X \in \mathcal{C}^*_m \right\}$$

where $\mathcal{C}^*_m$ is a cone of completely positive matrices, i.e. $X = YY'$ with Y an $m \times k$ matrix with $Y \geq 0$.

P. Amaral, I. Bomze, J. Júdice (2014). Copositivity and constrained fractional quadratic problems. *Mathematical Programming* 146, 325-350.

Conic Optimization



- In general working with $\mathcal{C}^*_m$ is difficult
- Usually, what is done is to work in a relaxation of this problem, replacing $\mathcal{C}^*_m$, by the cone of doubly nonnegative matrices $\mathcal{D}_m$

$$\varphi = max\left\{\mathbf{B} \cdot X : \mathbf{W} \cdot X = 1, X \in \mathcal{C}^*_m\right\}$$

$$\theta = max\left\{\mathbf{B} \cdot X : \mathbf{W} \cdot X = 1, X \in \mathcal{D}_m\right\}$$

In general $\varphi \leq \theta$. However, if $m \leq 4$ then $\varphi = \theta$.

Classification in two groups

Classification in two groups using the Mallows Distance

Considering two groups: C_1, C_2, an observation i and the respective quantile functions: $\overline{\Psi^{-1}_{D_{C1}}}(t)$, $\overline{\Psi^{-1}_{D_{C2}}}(t)$ and $\Psi^{-1}_{D(i)}(t)$

- The observation i is assigned to Group $C1$ if

$$D^2_M\left(\Psi^{-1}_{D(i)}(t), \overline{\Psi^{-1}_{D_{G1}}}(t)\right) < D^2_M\left(\Psi^{-1}_{D(i)}(t), \overline{\Psi^{-1}_{D_{G2}}}(t)\right)$$

- The observation i is assigned to Group $C2$ if

$$D^2_M\left(\Psi^{-1}_{D(i)}(t), \overline{\Psi^{-1}_{D_{G2}}}(t)\right) < D^2_M\left(\Psi^{-1}_{D(i)}(t), \overline{\Psi^{-1}_{D_{G1}}}(t)\right)$$

An observation i is assigned to the group for which the Mallows distance between its score and the score of the corresponding barycentric histogram is minimum.

USA 96 elections: Democrat/Republican state

Histogram-valued variables:
Pov: percentage of people under the poverty level;
Div: percentage of population who are divorced

- Only the states for which the number of records for all selected variables is higher than thirty, i.e. **twenty states** are considered.
- For all observations the subintervals of each histogram have the same weight (equiprobable) with frequency 0.20.

Groups:

Group 1 - Democrat: 12 States
Group 2 - Republican: 8 States

USA 96 elections: Democrat/Republican state

Discriminant function:

$$\Psi^{-1}_{D(i)}(t) \quad = 13.76\Psi^{-1}_{Pov(i)}(1-t) + 7.91\Psi^{-1}_{Div(i)}(t) + \overline{\Psi^{-1}_{S}}(t)$$

Parameters: Conic optimization - **Optimal solution**

Classification results: 80% well classified.

Outline

Concluding remarks

- From micro-data to macro-data:
  Interval and Distribution-valued data
- Take variability into account
- Several methodologies already developed
  for multivariate data analysis
- Histogram data : methods based on the Mallows distance between quantile functions
- New problems / challenges :
  distributions are not real numbers !

Concluding remarks

"Distributions are the numbers of the future"

(Berthold Schweizer, 1984)

Books and Main Papers

Books:

- Bock, H.-H., Diday, E. (2000): *Analysis of Symbolic Data: Exploratory methods for extracting statistical information from complex data.* Springer.
- Billard, L., Diday, E. (2007): *Symbolic Data Analysis: Conceptual Statistics and Data Mining.* Wiley.
- Diday, E., Noirhomme-Fraiture, M. (2008): *Symbolic Data Analysis and the SODAS Software.* Wiley.

Survey Papers:

- Billard, L., Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic Data Analysis. *JASA*, 98 (462), 470–487.
- Noirhomme-Fraiture, M., Brito, P. (2011). Far beyond the classical data models: Symbolic data analysis. *Statistical Analysis and Data Mining*, 4(2), 157–170.
- Brito, P. (2014). Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics. *WIREs Data Mining and Knowledge Discovery*, 4 (4), 281–295.


Congratulations!