Multivariate Analysis of Histogram Data

Paula Brito

FEP & LIAAD - INESC TEC, Universidade do Porto, Portugal
mpbrito@fep.up.pt ; www.fep.up.pt/docentes/mpbrito

Conservatoire National des Arts et Métiers
Paris, May 18th 2018

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.

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

Histogram-valued variables

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

B : set of probability or frequency distributions over a set of sub-intervals $\{I_{i1}, ..., I_{iK_i}\}$

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

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

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	$5, 8, 10, 12, 15, 20, 25, 25, 30, 32, 35, 35, 45, 52, 55, 60$

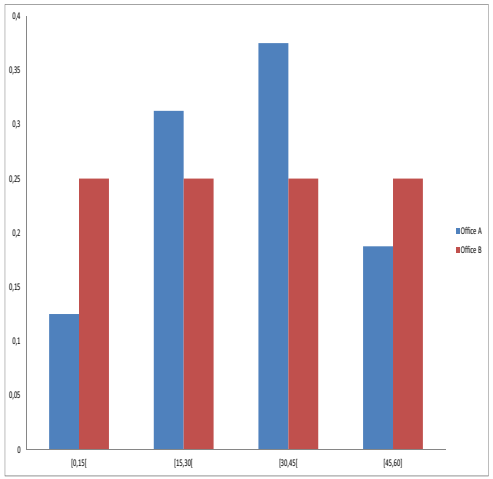
Average waiting time : 29.0 minutes for both offices

Description in terms of histograms :

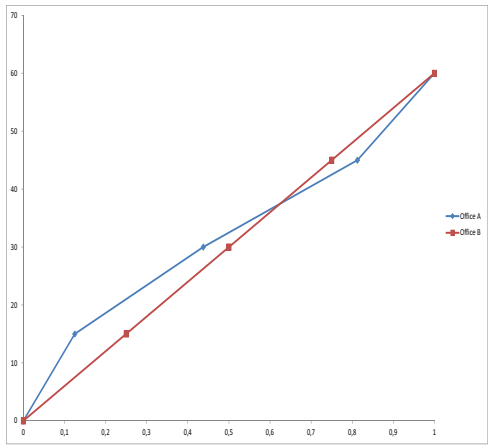
Office	Waiting Times (minutes)
A	$\{[0, 15[, 0.125; [15, 30[, 0.3125; [30, 45[, 0.375; [45, 60], 0.1875\}$
B	$\{[0, 15[, 0.25; [15, 30[, 0.25; [30, 45[, 0.25; [45, 60], 0.25\}$

Histogram-valued variables: Example

Histograms :



Quantile functions :



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

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

Descriptive Statistics for Histogram Variables

Assumming an Uniform distribution within each sub-interval of $Y_k(s_i)$, $i = 1, \ldots, n$, $I_{ik\ell} = [l_{ik\ell}, u_{ik\ell}]$, $\ell = 1, \ldots K_j$, $k = j, j'$ we have

- **Symbolic sample mean** :

$$\overline{Y_k} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{\ell=1}^{K_j} ((l_{ik\ell} + u_{ik\ell}) p_{ik\ell}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{K_j} (c_{ik\ell}\, p_{ik\ell})$$

- **Symbolic sample variance** :

$$S^2_{Y_k} == \frac{1}{3n} \sum_{i=1}^{n} \sum_{\ell=1}^{K_j} ((l^2_{ik} + l_{ik}\, u_{ik} + u^2_{ik}) p_{ik\ell}) - \overline{Y_k}^2$$

Billard and Diday (2003)

Descriptive Statistics for Histogram Variables

Covariance

Billard & Diday (2003), obtained from the empirical joint density function:

$$Cov(Y_j, Y_{j'}) = \frac{1}{4n} \sum_{i=1}^{n} \sum_{\ell=1}^{K_j} p_{ij\ell} p_{ij'\ell} (l_{ij} + u_{ij})(l_{ij'} + u_{ij'}) - \overline{Y_j}.\overline{Y_{j'}}$$

New definition in Billard (2008), considering a decomposition into Within observations Sum of Products (WithinSP) and Between observations Sum of Products (BetweenSP):

Descriptive Statistics for Histogram Variables

$$Cov(Y_j, Y_{j'}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{K_j} p_{ij\ell}p_{ij'\ell}\frac{(u_{ij}-l_{ij})(u_{ij'}-l_{ij'})}{12}}_{\text{WithinSP}} +$$

$$+\underbrace{\frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{K_j} p_{ij\ell}p_{ij'\ell}\left(\frac{l_{ij}+u_{ij}}{2}-\overline{Y_j}\right)\left(\frac{l_{ij'}+u_{ij'}}{2}-\overline{Y_{j'}}\right)}_{\text{BetweenSP}}$$

$$= \frac{1}{6n}\sum_{i=1}^{n}\sum_{\ell=1}^{K_j} p_{ij\ell}p_{ij'\ell}[2(l_{ij}-\overline{Y_j})(l_{ij'}-\overline{Y_{j'}}) + (l_{ij}-\overline{Y_j})(u_{ij'}-\overline{Y_{j'}}) + (u_{ij}-\overline{Y_j})(l_{ij'}-\overline{Y_{j'}}) + 2(u_{ij}-\overline{Y_j})(u_{ij'}-\overline{Y_{j'}})]$$

However, this Covariance function is not invariant under re-codifications of the histograms with different bins (Irpino 2013)

Histogram-valued variables: Distance measures

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

Divergency measures	
Kullback-Leibler	$D_{KL}(f,g) = \int_{\mathbb{R}} log\left(\frac{f(x)}{g(x)}\right) f(x)dx$
Jeffrey	$D_J(f,g) = D_{KL}(f,g) + D_{KL}(g,f)$
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	$D_{var}(f,g) = \int_{\mathbb{R}} \lvert f(x) - g(x)\rvert dx$
Wasserstein	$D_W(f,g) = \int_{\mathbb{R}} \lvert F^{-1}(x) - G^{-1}(x)\rvert dx$
Mallows	$D_M(f,g) = \sqrt{\int_0^1 \left(F^{-1}(x) - G^{-1}(x)\right)^2 dx}$
Kolmogorov	$D_K(f,g) = \max_{\mathbb{R}} \lvert F(x) - G(x)\rvert$

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 :

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

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

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

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

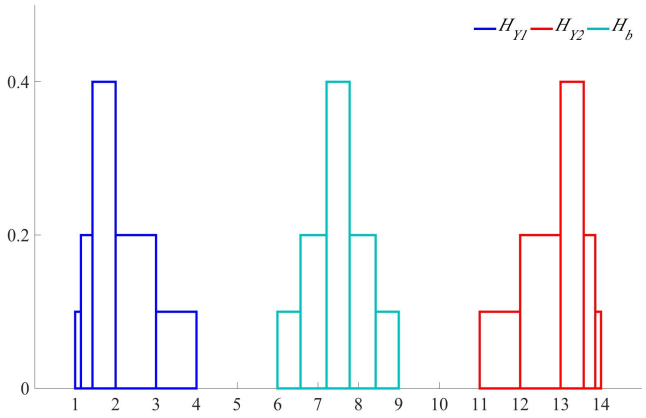
Need to re-write the histograms - and quantile functions - with the same weight decomposition

Descriptive Statistics for Histogram Variables: Barycenter

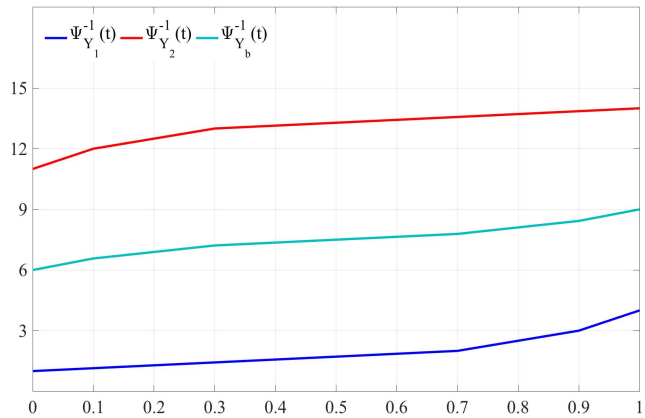
$H_1 = \{[1; 2[; 0.7; [2; 3[; 0.2; [3; 4]; 0.1\}$
$H_2 = \{[11; 12[; 0.1; [12; 13[; 0.2; [13; 14]; 0.7\}$

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 fulfills 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

Outline

Clustering

Methods based directly on dissimilarity measures : straightforward adaptation

- Hierarchical clustering : Complete linkage, Single linkage
- Non-hierarchical clustering : k-Medoïds

The Huyghens decomposition of the Mallows distance allows for the extension of other clustering methods :

- Ward hierarchical clustering (Irpino and Verde, 2006)
- Divisive clustering (Brito and Chavent, 2011)
- Non-hierarchical dynamical clustering (k-Means like) (Irpino and Verde, 2006)

Divisive Clustering

Joint work with Marie Chavent (IMB & INRIA CQFD, Univ. Bordeaux 2, France)

- Divisive clustering algorithms proceed top-down
- Starting with S, the set to be clustered
- Performing a bipartition of one cluster at each step
- At step m a partition of S in m clusters is present
- One will be further divided in two sub-clusters
- The cluster to be divided and the splitting rule chosen to obtain a partition in $m+1$ clusters minimizing intra-cluster dispersion

Brito, P., Chavent, M. (2012). Divisive Monothetic Clustering for Interval and Histogram-Valued Data. In: Proc. ICPRAM 2012 - 1st International Conference on Pattern Recognition Applications and Methods, Vilamoura, Portugal.

The criterion

"Quality" of a partition $P_m = \left\{ C_1^{(m)}, C_2^{(m)}, \ldots, C_m^{(m)} \right\}$ measured by the sum of intra-cluster dispersion for each cluster :

$$Q(m) = \sum_{\alpha=1}^{K} I(C_\alpha) = \sum_{\alpha=1}^{K} \sum_{s_i, s_{i'} \in C_\alpha^{(m)}} D^2(s_i, s_{i'})$$

$$D^2(s_i, s_{i'}) = \sum_{j=1}^{p} d^2(x_{ij}, x_{i'j})$$

d : quadratic distance between distributions

At each step :
one cluster is chosen to be split in two sub-clusters
$Q(m+1)$ is minimized ($Q(m) - Q(m+1)$ maximized)

Binary questions

Bipartition to be performed at each step :
defined by one single variable considering conditions of the type :

$$R_{j\ell} := Y_j \leq \bar{I}_{j\ell}, \ell = 1, \ldots, K_j - 1, j = 1, \ldots, p$$

$R_{j\ell}$: bipartition of a cluster :

sub-cluster 1 : elements who verify condition $R_{j\ell} : Y_j \leq \bar{I}_{j\ell}$
sub-cluster 2 : those who do not : $Y_j > \bar{I}_{j\ell}$

Assignement

$s_i \in S$ verifies the condition $R_{j\ell} : Y_j \leq \bar{I}_{j\ell}$ iff $\sum_{\alpha=1}^{\ell} p_{ij\alpha} \geq 0.5$

The sequence of conditions :
necessary and sufficient condition for cluster membership

The obtained clustering is **monothetic** :
each cluster is represented by a conjunction of properties in the descriptive variables

Binary questions and assignement: example

	Age	Marks
Class 1	([10, 11[, 0.50; [11, 12[, 0.50; [12, 14], 0.00)	([5, 10[, 0.2; [10, 12[, 0.5; [12, 14[, 0.133; [14, 15[, 0.067; [15, 16[, 0.033; [16, 18[, 0.067; [18, 19], 0.0)
Class 2	([10, 11[, 0.00; [11, 12[, 0.33; [12, 14], 0.67)	([5, 10[, 0.05; [10, 12[, 0.3; [12, 14[, 0.25; [14, 15[, 0.1; [15, 16[, 0.1; [16, 18[, 0.133; [18, 19], 0.067)

First step - binary questions :

$Age \leq 11$, $Age \leq 12$
$Marks \leq 10$, $Marks \leq 12$, $Marks \leq 14$, $Marks \leq 15$, $Marks \leq 16$, $Marks \leq 18$

If condition $Age \leq 12$ is selected :
sub-cluster 1 contains Class 1 and is described by "$Age \leq 12$"
sub-cluster 2 contains Class 2 and is described by "$Age > 12$"

Algorithm

- Initialization : $P_1 = \{C_1^{(1)} \equiv S\}$
- $P_m = \{C_1^{(m)}, \ldots, C_m^{(m)}\}$: current partition at step m
  Determine the cluster $C_M^{(m)}$ and the binary question $R_{j\ell} := Y_j \leq \bar{I}_{j\ell}$:
  new partition $P_{m+1} = \{C_1^{(m+1)}, \ldots, C_{m+1}^{(m+1)}\}$ minimizes
  $$Q(m+1) = \sum_{\ell=1}^{m+1} \sum_{s_i, s_{i'} \in C_\ell^{(m+1)}} D^2(s_i, s_{i'})$$
  among partitions in $m+1$ clusters obtained by splitting a cluster of P_m in two clusters
- Minimize Q(m+1) : equivalent to maximize
  $\Delta Q = I(C_M^{(m)}) - (I(C_1^{(m+1)}) + I(C_2^{(m+1)}))$
- Fixed number of clusters K is attained
  or P has n clusters, each with a single element (step n):
  algorithm stops

Application: Social and crime data in USA states

State	Cities	X_1	X_2	...
Alabama	Selma	2.73	5.84	...
	Bessemer	2.66	5.83	...
	⋮	⋮	⋮	⋮
	Madison	2.65	1.34	...
California	San Pablo	2.61	4.19	...
	Glendale	2.49	2.34	...
	⋮	⋮	⋮	⋮
	West Sacramento	2.75	3.45	...
Florida	Rockledge	3.61	5.38	...
	Ormond Beach	3.69	4.04	...
	⋮	⋮	⋮	⋮
	Fort Myers	2.65	5.78	...
	⋮	⋮	⋮	⋮

Social and crime data in USA states

Aggregated histogram data :

State	**pctEmploy**
Alabama	$[42.6, 50.7[, 0.2; [50.7, 55.1[, 0.2; [55.1, 59.0[, 0.2;$ $[59.0, 62.2[, 0.2; [62.2, 77.0], 0.2$
California	$[46.0, 51.8[, 0.2; [51.4, 56.2[, 0.2; [56.2, 61.6[, 0.2;$ $[61.6, 67.7[, 0.2; [67.7, 76.2], 0.2$
...	...

Social and crime data in USA states: variables

CRIMES

- murdPerPop: number of murders per 100K population
- robbbPerPop: number of robberies per 100K population
- assaultPerPop: number of assaults per 100K population
- burglPerPop: number of burglaries per 100K population
- larcPerPop: number of larcenies per 100K population
- autoTheftPerPop: number of auto thefts per 100K population
- arsonsPerPop: number of arsons per 100K population

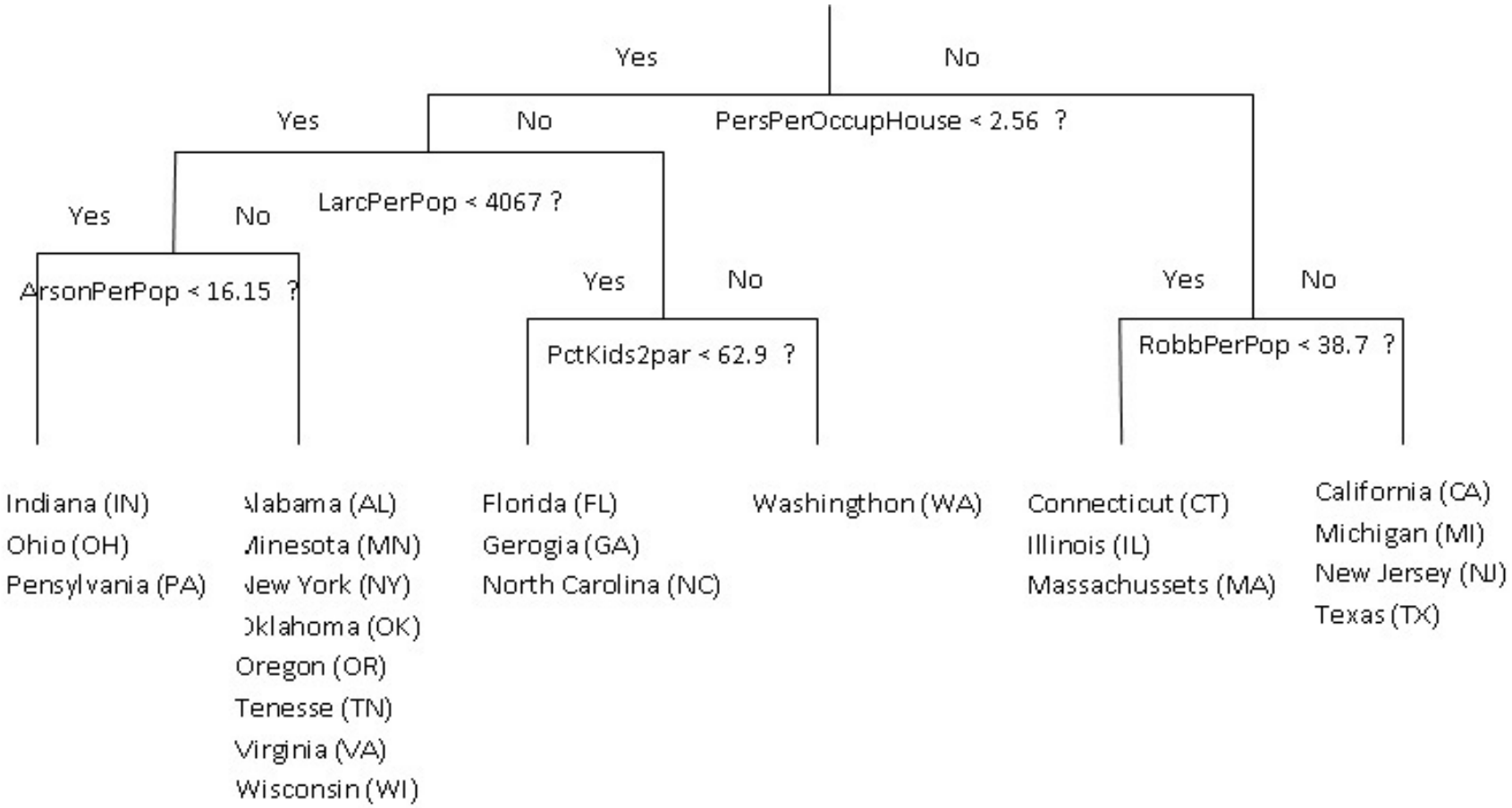
SOCIAL

- perCapInc: per capita income
- PctPopUnderPov: percentage of people under the poverty level
- PersPerOccupHous: mean persons per household
- PctKids2Par: percentage of kids in family housing with two parents
- PctVacantBoarded: percent of vacant housing that is boarded up
- NumKindsDrugsSeiz: number of different kinds of drugs seized
- LemasTotReqPerPop: total requests for police per 100K popuation

Application: Social and crime data in USA states

- Data gathered for 2216 USA cities, aggregated by state - 22 states retained
- 14 numerical variables - distributions represented by histogram-valued variables
- Sub-intervals defined by $k \times 20\%$ quantiles
- Mallows distance between distributions for each state
- Partition into six clusters

Social and crime data in USA states: the dendrogram



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 lies 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 function, the linear combination includes:

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

Linear combination between quantile functions

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

$$\Psi_Y^{-1}(t) = \gamma + \sum_{j=1}^{p} a_j \Psi_{X_j}^{-1}(t) - \sum_{j=1}^{p} b_j \Psi_{X_j}^{-1}(1-t)$$

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

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:

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

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

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

Distribution and Symmetric Distribution Linear Regression model

- Experiments, both with small real data sets and simulated data: the model appears to work well
- The goodness-of-fit measure shows good behavior

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

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
- 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

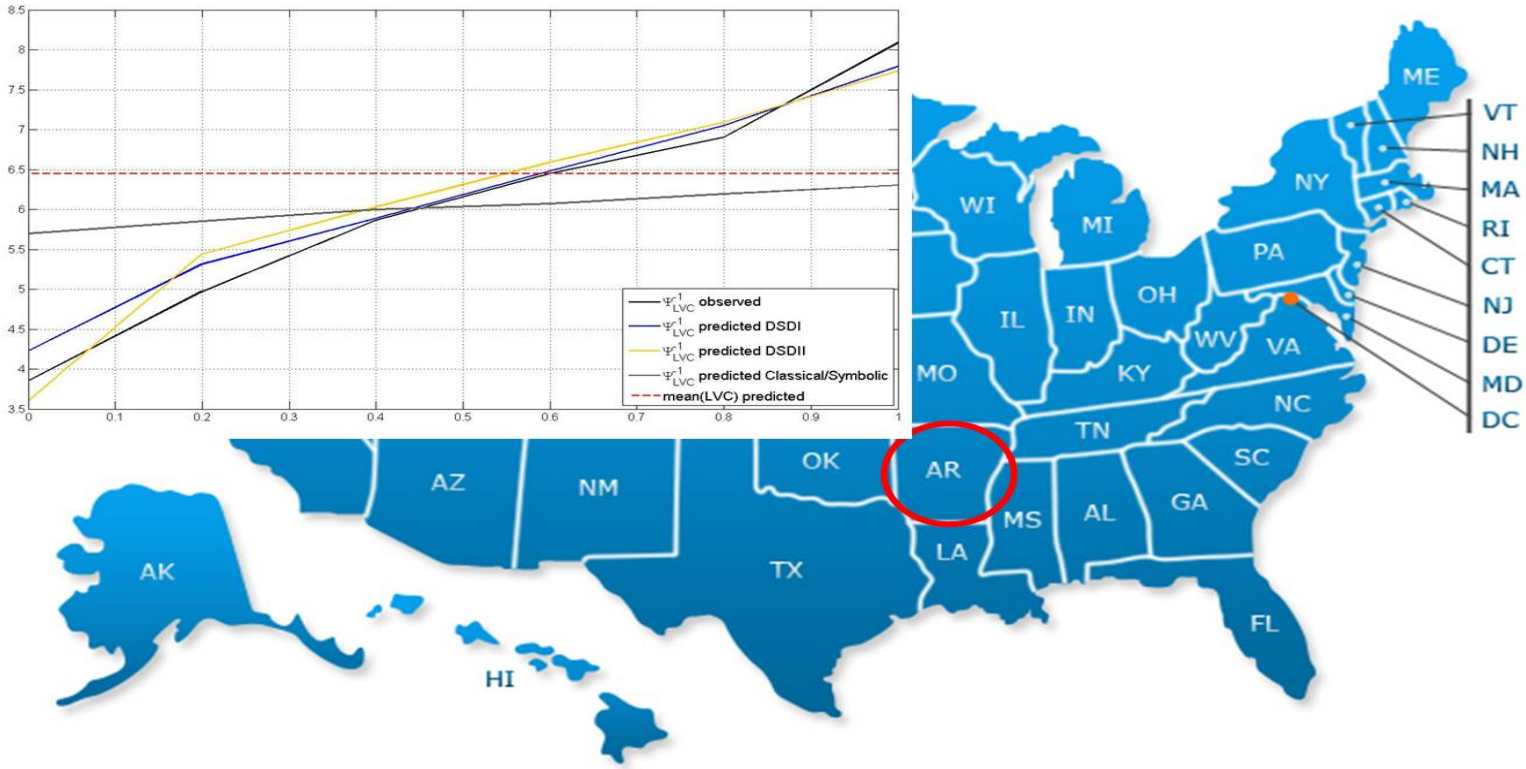
Model DSD I:

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

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.

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

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

Linear Discriminant Analysis

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

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

Given the Huyghens decomposition,

$$\sum_{i=1}^{n} D^2_M(\Psi^{-1}_{S(i)}(t), \overline{\Psi^{-1}_{S}}(t)) =$$

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

we may write

$$\gamma' T \gamma = \gamma' B \gamma + \gamma' W \gamma$$

Linear Discriminant Analysis

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

Optimization problem:

Maximize the ratio $\lambda = \frac{\gamma' \mathbf{B} \gamma}{\gamma' \mathbf{W} \gamma}$ subject to $\gamma \geq 0$

Fractional quadratic problem with linear constraints

- Hard optimization problem
- Nonconvex
- Easy to find a good solution
- Difficult to prove optimality

Using BARON - Branch and Bound, and Conic Optimization
For linear discriminant situations with two variables (in $\mathcal{C}_4^*$) it is possible to prove that the good solution is an optimal solution

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 USA states

Goal: Classify the USA states as Democrat or Republican considering socio-economic caracteristics

Original data - Microdata: Socio-economic data from the '90 Census Presidential Election results of 1996
First level units: Cities of the USA states
Observations associated with each unit: The records (real values) associated with the cities in USA states.

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

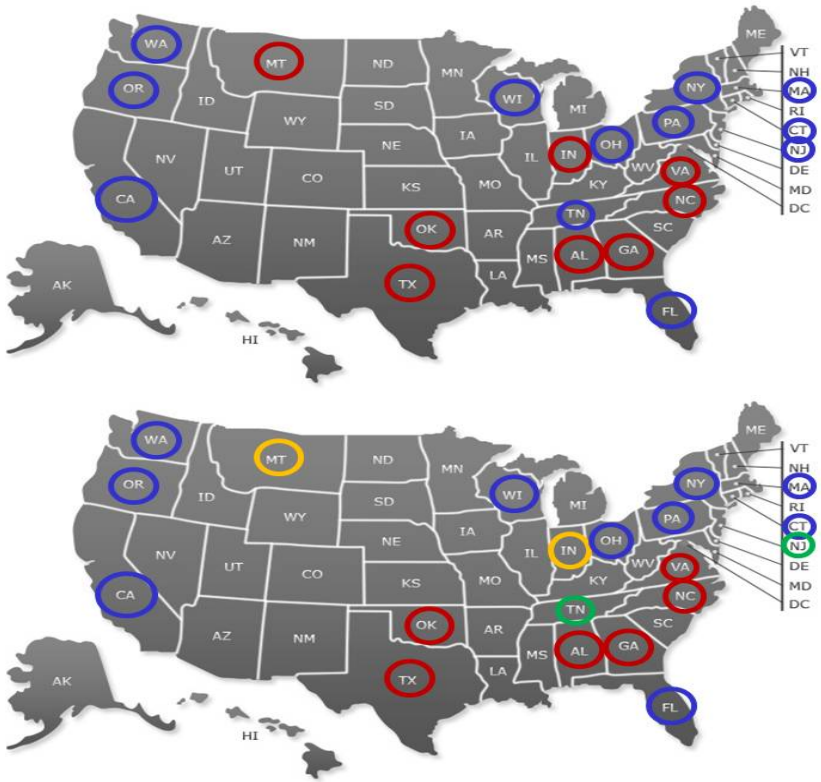
Discriminant function:

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

Parameters: Conic optimization - **Optimal solution**

Classification results: 80% well classified.

Democrat/Republican USA states



Democrat and Republican states

Yellow - Republican states predicted as Democrat

Green - Democrat states predicted as Republican

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 !

Books and Main Papers


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.


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.