Symbolic Data Analysis
Hands-on Session

4th International Summer School on Data Science

September 13, 2019

Table of contents

Introduction to symbolic data

Introduction to symbolic data

Types of data:

- Classical data value X: single point in p-dimensional space
- Symbolic data value Z: **hypercube** or **Cartesian product of distributions** in $\mathbb{R}^p$

Introduction to symbolic data

Types of data:

- Classical data value X: single point in p-dimensional space
- Symbolic data value Z: **hypercube** or **Cartesian product of distributions** in $\mathbb{R}^p$

How do symbolic data arise:

1. Aggregate data (e.g. research interest: classes or groups)
   - Age $\times$ gender categories
   - Pileus cap width (*arorae*) $= [3.0, 8.0]$ (*a* mushroom/*the* mushroom)

Introduction to symbolic data

Types of data:

- Classical data value X: single point in p-dimensional space
- Symbolic data value Z: **hypercube** or **Cartesian product of distributions** in $\mathbb{R}^p$

How do symbolic data arise:

1. Aggregate data (e.g. research interest: classes or groups)
   - Age $\times$ gender categories
   - Pileus cap width (*arorae*) $= [3.0, 8.0]$ (*a* mushroom/*the* mushroom)
2. Naturally occurring symbolic data
   - Pulse data – recorded in a range (e.g. 64 ± 2)
   - Birds' colors (e.g. $\{\text{black}\}$, $\{\text{yellow}, \text{red}\}$, $\{\text{yellow}, \text{blue}\}$)

Types of symbolic data

We can define several different types of symbolic data

Types of symbolic data

We can define several different types of symbolic data

- **Multi-valued symbolic random variable** – takes one or more values from the list of values in its domain
  - Colors associated with birds

$$\text{color of magpie} = \{\text{black}, \text{white}\}$$
$$\text{color of cardinal} = \{\text{red}, \text{black}\}$$

Types of symbolic data

We can define several different types of symbolic data

- **Multi-valued symbolic random variable** – takes one or more values from the list of values in its domain
  - Colors associated with birds

$$\text{color of magpie} = \{\text{black}, \text{white}\}$$
$$\text{color of cardinal} = \{\text{red}, \text{black}\}$$

- Interval-valued symbolic random variable – takes value in an interval

Types of symbolic data

We can define several different types of symbolic data

- **Multi-valued symbolic random variable** – takes one or more values from the list of values in its domain
  - Colors associated with birds

$$\text{color of magpie} = \{\text{black}, \text{white}\}$$
$$\text{color of cardinal} = \{\text{red}, \text{black}\}$$

- Interval-valued symbolic random variable – takes value in an interval
- Histogram valued symbolic random variable – takes value on non-overlapping intervals with a weight assigned to each particular interval

Interval-valued symbolic variable

An **interval-valued** symbolic random variable Z is one that takes values in an interval, i.e. $Z = [a, b] \subset \mathbb{R}$, with $a \leq b$, $a, b \in \mathbb{R}$.

- We will primarily focus on this type during the session (by aggregating the Iris dataset)

Creating symbolic data

We will use the Iris dataset that quantifies the morphologic variation of iris flowers of three different species: setosa, versicolor, and virginica.

Creating symbolic data

We will use the Iris dataset that quantifies the morphologic variation of iris flowers of three different species: setosa, versicolor, and virginica.

Assignment 1

Get a glimpse of the Iris dataset and display the first rows.

Creating symbolic data

Suppose we aggregate the variables in the Iris dataset by species in order to obtain an interval-valued symbolic dataset.

Creating symbolic data

Suppose we aggregate the variables in the Iris dataset by species in order to obtain an interval-valued symbolic dataset.

Assignment 2

Find minimal and maximal values of each feature in the Iris dataset for each species. This will allow us to create interval-valued symbolic variable as $[x_{\text{min}}, x_{\text{max}}]$.

Creating symbolic data

	Sepal length	Sepal width	Petal length	Petal width
Setosa	$[4.3, 5.8]$	$[2.3, 4.4]$	$[1.0, 1.9]$	$[0.1, 0.6]$
Versicolor	$[4.9, 7.0]$	$[2.0, 3.4]$	$[3.0, 5.1]$	$[1.0, 1.8]$
Virginica	$[4.9, 7.9]$	$[2.2, 3.8]$	$[4.5, 6.9]$	$[1.4, 2.5]$

Creating symbolic data

	Sepal length	Sepal width	Petal length	Petal width
Setosa	$[4.3, 5.8]$	$[2.3, 4.4]$	$[1.0, 1.9]$	$[0.1, 0.6]$
Versicolor	$[4.9, 7.0]$	$[2.0, 3.4]$	$[3.0, 5.1]$	$[1.0, 1.8]$
Virginica	$[4.9, 7.9]$	$[2.2, 3.8]$	$[4.5, 6.9]$	$[1.4, 2.5]$

Assignment 3

Import `RSDA` package and create a symbolic dataset from the Iris data. Use `classic.to.sym` function.

Univariate descriptive statistics

Basic univariate descriptive statistics

Basic descriptive statistics for one random variable include mean and variance. We will focus on their symbolic data analogues, specifically for interval-valued variables.

Note: since a symbolic variable $\xi = [a, a]$ is equivalent to its classical counterpart $x = a$, all descriptive statistic for ξ and x will have same values.

Symbolic sample mean

For an interval-valued random variable Z, the **symbolic sample mean** is given by

$$\bar{Z} = \frac{1}{m} \sum_{u \in E} \frac{b_u + a_u}{2},$$

where $u \in E$ represents an observation of Z.

Empirical density function

In order to formally derive the symbolic sample mean, we can use the empirical density function of an interval variable

$$f(\xi) = \frac{1}{m} \sum_{u:\xi \in Z(u)} \left(\frac{1}{b_u - a_u} \right).$$

Empirical density function

In order to formally derive the symbolic sample mean, we can use the empirical density function of an interval variable

$$f(\xi) = \frac{1}{m} \sum_{u:\xi \in Z(u)} \left(\frac{1}{b_u - a_u} \right).$$

Note that the summation is only over those observations u for which $\xi \in [a_u, b_u] = Z(u)$. We also used the assumption that individual classical values are uniformly distributed over the interval $Z(u)$.

Empirical density function

In order to formally derive the symbolic sample mean, we can use the empirical density function of an interval variable

$$f(\xi) = \frac{1}{m} \sum_{u:\xi \in Z(u)} \left(\frac{1}{b_u - a_u} \right).$$

Note that the summation is only over those observations u for which $\xi \in [a_u, b_u] = Z(u)$. We also used the assumption that individual classical values are uniformly distributed over the interval $Z(u)$.

The symbolic sample mean follows from the expectation

$$\bar{Z} = \int_{-\infty}^{\infty} \xi f(\xi) d\xi.$$

Symbolic sample mean – Iris dataset example

	Setosa	Versicolor	Virginica
Sepal length	$[4.3, 5.8]$	$[4.9, 7.0]$	$[4.9, 7.9]$

For Sepal length variable of the Iris dataset we have:

$$\bar{Z} = \frac{1}{3}\left(\frac{4.3 + 5.8}{2} + \frac{4.9 + 7.0}{2} + \frac{4.9 + 7.9}{2}\right) = 5.8.$$

Symbolic sample mean – Iris dataset example

	Setosa	Versicolor	Virginica
Sepal length	$[4.3, 5.8]$	$[4.9, 7.0]$	$[4.9, 7.9]$

For Sepal length variable of the Iris dataset we have:

$$\bar{Z} = \frac{1}{3}\left(\frac{4.3+5.8}{2} + \frac{4.9+7.0}{2} + \frac{4.9+7.9}{2}\right) = 5.8.$$

Assignment 4

Calculate the symbolic sample mean for all variables in the Iris dataset. Use `sym.mean` function.

Symbolic sample variance

For an interval-valued random variable Z, the **symbolic sample variance** is given by

$$
\begin{aligned}
S^2 &= \frac{1}{3m} \sum_{u \in E} \left(b_u^2 + a_u b_u + a_u^2 \right) - \bar{Z}^2 \\
&= \frac{1}{3m} \sum_{u \in E} \left(b_u^2 + a_u b_u + a_u^2 \right) - \frac{1}{4m^2} \left[\sum_{u \in E} \left(b_u + a_u \right) \right]^2 .
\end{aligned}
$$

Symbolic sample variance

Similarly, we can verify the symbolic sample variance equation using

$$
\begin{aligned}
S^2 &= \int_{-\infty}^{\infty} \left(\xi - \bar{Z}\right)^2 f(\xi) d\xi \\
&= \int_{-\infty}^{\infty} \xi^2 f(\xi) d\xi - \bar{Z}^2.
\end{aligned}
$$

Symbolic sample variance – Iris dataset example

	Setosa	Versicolor	Virginica
Sepal length	$[4.3, 5.8]$	$[4.9, 7.0]$	$[4.9, 7.9]$

For Sepal length variable of the Iris dataset we have:

$$
\begin{aligned}
S^2 = \frac{1}{9}\Big[&\left(4.3^2 + 4.3 \cdot 5.8 + 5.8^2\right) \\
&+ \left(4.9^2 + 4.9 \cdot 7.0 + 7.0^2\right) \\
&+ \left(4.9^2 + 4.9 \cdot 7.9 + 7.9^2\right) \Big] - 5.8^2 = 0.75.
\end{aligned}
$$

Symbolic sample variance – Iris dataset example

	Setosa	Versicolor	Virginica
Sepal length	$[4.3, 5.8]$	$[4.9, 7.0]$	$[4.9, 7.9]$

For Sepal length variable of the Iris dataset we have:

$$\begin{aligned} S^2 = \frac{1}{9}\Big[&\left(4.3^2 + 4.3 \cdot 5.8 + 5.8^2\right) \\ &+ \left(4.9^2 + 4.9 \cdot 7.0 + 7.0^2\right) \\ &+ \left(4.9^2 + 4.9 \cdot 7.9 + 7.9^2\right) \Big] - 5.8^2 = 0.75. \end{aligned}$$

Assignment 5

Calculate the symbolic sample variance for all variables in the Iris dataset. Use `sym.variance` function.

Multivariate descriptive statistics

Multivariate descriptive statistics

We will focus on basic multivariate descriptive statistics of covariance and correlation. Similarly as before, we will expect equivalence to classical covariance and correlation for symbolic variables $\xi = [a, a]$.

Empirical covariance function

For interval valued variables Z_1 and Z_2, the **empirical covariance function** $\mathrm{Cov}\,(Z_1, Z_2)$ is given by

$$\mathrm{Cov}\,(Z_1, Z_2) = \frac{1}{3m} \sum_{u \in E} G_1 G_2 \left[Q_1 Q_2 \right]^2 ,$$

where, for $j = 1, 2$,

$$Q_j = \left(a_{uj} - \bar{Z}_j\right)^2 + \left(a_{uj} - \bar{Z}_j\right)\left(b_{uj} - \bar{Z}_j\right) + \left(b_{uj} - \bar{Z}_j\right)^2$$
$$G_j = \begin{cases} -1, & \text{if } \bar{Z}_{uj} \leq \bar{Z}_j \\ +1, & \text{if } \bar{Z}_{uj} > \bar{Z}_j \end{cases} .$$

$\bar{Z}_j$ is the symbolic sample mean, and $\bar{Z}_{uj} = (a_{uj} + b_{uj})/2$.

Empirical covariance function

For $Z_1 = Z_2 = Z$ we have

$$\begin{aligned}
\mathrm{Cov}\,(Z,Z) &= \frac{1}{3m}\sum_{u\in E}(a_u-\bar{Z})^2+(a_u-\bar{Z})(b_u-\bar{Z})+(b_u-\bar{Z})^2 \\
&= \frac{1}{3m}\sum_{u\in E} a_u^2-2a_u\bar{Z}+\bar{Z}^2+a_ub_u-a_u\bar{Z}-b_u\bar{Z}+\bar{Z}^2 \\
&\qquad +b_u^2-2b_u\bar{Z}+\bar{Z}^2 \\
&= \frac{1}{3m}\sum_{u\in E}(a_u^2+a_ub_u+b_u^2)+\frac{1}{3m}\sum_{u\in E}3\bar{Z}^2-\frac{1}{3m}\sum_{u\in E}3\bar{Z}(a_u+b_u) \\
&= \frac{1}{3m}\sum_{u\in E}(a_u^2+a_ub_u+b_u^2)+\bar{Z}^2-\bar{Z}\cdot\underbrace{\frac{1}{m}\sum_{u\in E}(a_u+b_u)}_{2\bar{Z}} \\
&= \frac{1}{3m}\sum_{u\in E}(a_u^2+a_ub_u+b_u^2)-\bar{Z}^2 \\
&= S^2.
\end{aligned}$$

Empirical covariance function – Iris dataset example

	Setosa	Versicolor	Virginica
Sepal length	$[4.3, 5.8]$	$[4.9, 7.0]$	$[4.9, 7.9]$
Sepal width	$[2.3, 4.4]$	$[2.0, 3.4]$	$[2.2, 3.8]$

Suppose we want to calculate covariance between Sepal length and Sepal width variables. Mean values are $\bar{Z} = 5.8$ and $\bar{Z} = 3.0167 \approx 3.0$. For observation setosa we have

$$Q_1 = (4.3 - 5.8)^2 + (4.3 - 5.8)(5.8 - 5.8) + (5.8 - 5.8)^2 = 2.25$$
$$Q_2 = (2.3 - 3.0)^2 + (2.3 - 3.0)(4.4 - 3.0) + (4.4 - 3.0)^2 = 1.44.$$

$G_1 = -1$ since $\bar{Z}_{11} = (4.3 + 5.8)/2 = 5.05 \leq \bar{Z}_1 = 5.8$, and $G_2 = -1$ since $\bar{Z}_{12} = (2.3 + 4.4)/2 = 3.35 > \bar{Z}_2 = 3.0167$.

Empirical covariance function – Iris dataset example

By repeating the procedure for other two observations, versicolor and virginica, we have

$$
\begin{aligned}
\operatorname{Cov}(Z_1, Z_2) = \frac{1}{9}\Big[& (-1) \cdot 1 \cdot \sqrt{2.25 \cdot 1.44} \\
& + 1 \cdot (-1) \cdot \sqrt{1.17 \cdot 0.79} \\
& + 1 \cdot (-1) \cdot \sqrt{3.33 \cdot 0.64}\Big] \\
= -0.46890.
\end{aligned}
$$

Empirical covariance function – Iris dataset example

By repeating the procedure for other two observations, versicolor and virginica, we have

$$
\begin{aligned}
\operatorname{Cov}(Z_1, Z_2) = \frac{1}{9}\Big[& (-1) \cdot 1 \cdot \sqrt{2.25 \cdot 1.44} \\
& + 1 \cdot (-1) \cdot \sqrt{1.17 \cdot 0.79} \\
& + 1 \cdot (-1) \cdot \sqrt{3.33 \cdot 0.64}\Big] \\
= -0.46890. &
\end{aligned}
$$

Assignment 6

Calculate the empirical covariance between sepal length and sepal width variables of the Iris dataset. Use `sym.cov` function.

Empirical correlation function

For interval-valued variables Z_1 and Z_2, the **empirical correlation coefficient** $r(Z_1, Z_2)$ is given by

$$r(Z_1, Z_2) = \frac{\operatorname{Cov}(Z_1, Z_2)}{\sqrt{S_{Z_1}^2 S_{Z_2}^2}},$$

where $S_{Z_1}^2$ and $S_{Z_2}^2$ represent the symbolic sample variance of Z_1 and Z_2, respectively.

Empirical correlation function – Iris dataset example

	Setosa	Versicolor	Virginica
Sepal length	$[4.3, 5.8]$	$[4.9, 7.0]$	$[4.9, 7.9]$
Sepal width	$[2.3, 4.4]$	$[2.0, 3.4]$	$[2.2, 3.8]$

Suppose we want to calculate correlation between Sepal length and Sepal width variables. We can use the previous result, $\mathrm{Cov}\,(Z_1, Z_2) = -0.46890$, and with variances $S^2_{Z_1} = 0.75$ and $S^2_{Z_2} = 0.31861$ we have

$$r(Z_1, Z_2) = \frac{-0.46890}{\sqrt{0.75 \cdot 0.31861}} = -0.95923.$$

Empirical correlation function – Iris dataset example

Assignment 7

Calculate the empirical correlation between Sepal length and Sepal width variables of the Iris dataset. Use `sym.cor` function.

Principal component analysis for symbolic data

Principal component analysis for symbolic data

Recall that the principal component analysis is a method designed to reduce p-dimensional observations into s-dimensional components.

We will consider two methods of conducting PCA on symbolic data:

- Vertices method
- Centers method

Example data

We will use the blood pressure interval-valued dataset `lynne1` from the `RSDA` package. The dataset has three interval-valued variables: pulse rate, systolic pressure, and diastolic pressure.

Example data

We will use the blood pressure interval-valued dataset `lynne1` from the `RSDA` package. The dataset has three interval-valued variables: pulse rate, systolic pressure, and diastolic pressure.

Assignment 8

Import blood pressure interval-valued data (`lynne1`) and plot the data using `sym.scatterplot`.

Vertices method

Data representation:

- Each symbolic variable for a given object is represented by a hyper-rectangle with 2^p vertices.
- Object is represented by a $2^p \times p$ matrix M_u, containing the coordinate values for the hyper-rectangle.
- As this is done for each object, a $(m \cdot 2^p \times p)$ matrix M is constructed as follows:

$$M = \begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ M_m \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1p} \\ & \ddots & \\ b_{11} & \cdots & b_{1p} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} a_{m1} & \cdots & a_{mp} \\ & \ddots & \\ b_{m1} & \cdots & b_{mp} \end{bmatrix} \end{pmatrix}$$

Vertices method

For example, if there are two variables, $p = 2$, the data $\xi_u = ([a_{u1}, b_{u1}], [a_{u2}, b_{u2}])$ is transformed to the $2^2 \times 2$ matrix:

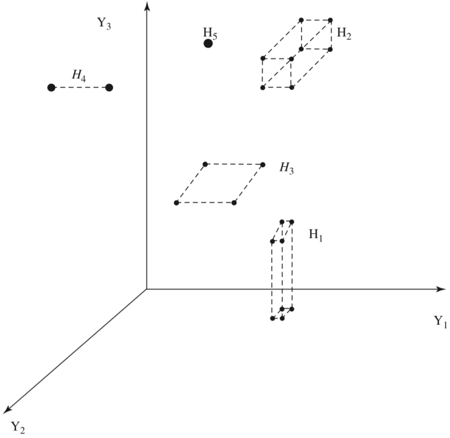
$$M_u = \begin{bmatrix} a_{u1} & a_{u2} \\ a_{u1} & b_{u2} \\ b_{u1} & a_{u2} \\ b_{u1} & b_{u2} \end{bmatrix}$$

and likewise for M.

The matrix M is now treated as though it represents classical data for $n = m \cdot 2^p$ individuals. Therefore, a classical PCA can be applied.

Vertices method - Geometrical interpretation

First figure shows different types of hyperrectangles H_u that can be represented by the matrix M_u, where each row contains values of each vertex.



Vertices method – Geometrical interpretation

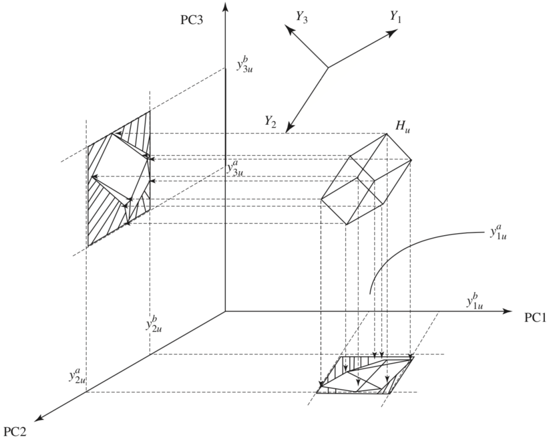
The `RSDA` includes a function for plotting 3D hyperrectangels `sym.scatterplot3d`.

Assignment 9

Plot the blood pressure dataset using `sym.scatterplot3d`.

Vertices method - Geometrical interpretation

Second figure shows a 3-dimensional hyper-rectangle H_u and its projections onto the first and second principal component plane, and onto the second and third principal component plane.



The projection is the maximum covering area rectangle (MCAR).

Vertices method – Blood pressure dataset example

Assignment 10

Apply PCA vertices method to the blood pressure dataset. Use `sym.interval.pca` function.

Centers method

We can also define a different approach – instead of using the vertices of hyper-rectangles, it is possible to use their centers.

In this case, each object $\xi_u = ([a_{u1}, b_{u1}], \dots, [a_{up}, b_{up}])$ is transformed to

$$x_u^c = (x_{u1}^c, \dots, x_{up}^c),\ u = 1, 2, ..., m,$$

where

$$x_{uj}^c = \frac{a_{uj} + b_{uj}}{2},\ j = 1, 2, ..., p$$

.

The symbolic data matrix X is transformed to a classical $m \times p$ matrix X^c with classical variables $x_1^c, x_2^c, \dots x_p^c$.
Then, the classical PCA is applied to X^c.

Centers method – Blood pressure dataset example

Assignment 11

Apply PCA centers method to the blood pressure dataset. Use `sym.interval.pca` function.

Thank you!