What if our data were intervals?

Lina Oliveira & M. Rosário Oliveira

CAMGSD, CEMAT, and Department of Mathematics, Instituto Superior Técnico

MMAC Day, March 14, 2024

Joint with R. Girão Serrão and M.Rosário Oliveira

"Theoretical derivation of interval principal component analysis",

Information Sciences, 2023

Conventional data/**Symbolic Data**:

- Each **object** is characterised by several **variables**
- Data is organized as an $n \times p$ matrix: rows correspond to **objects**, columns to **variables**

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \qquad \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

For example,

x_{11} = real number $\qquad$ x_{11} = interval $[a, b]$ of real numbers

Conventional data/Symbolic Data:

Data matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

random vector $\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \rightarrow \mathbb{R}^p$

Data matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

random vector $\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \rightarrow \mathbb{IR}^p$

Random vector

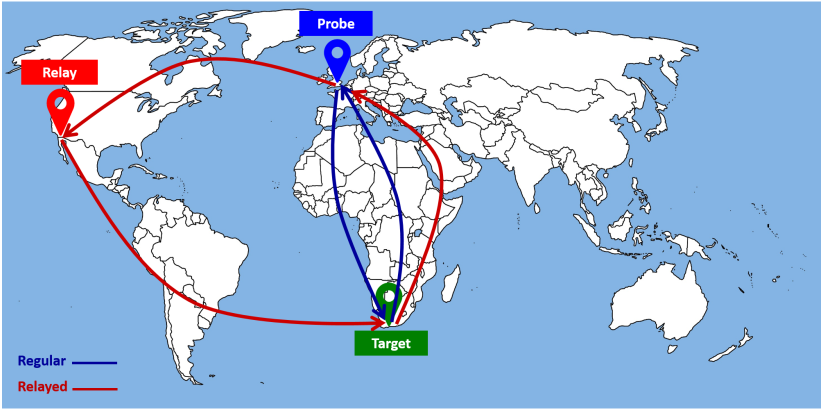
$[x_{11} \cdots x_{1p}]^T \in \mathbb{R}^p$ is a realisation of $\boldsymbol{X}$

Interval-valued random vector

$[x_{11} \cdots x_{1p}]^T \in \mathbb{IR}^p$ is a realisation of $\boldsymbol{X}$

WHY INTERVALS?

Detecting Internet traffic redirection attacks



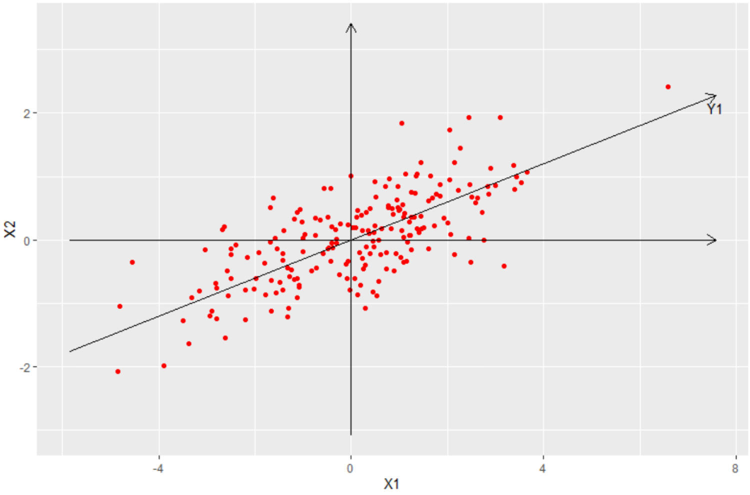
in A. Subtil, M.R. Oliveira, R. Valadas, A. Pacheco, and P. Salvador "Internet-Scale Traffic Redirection Attacks Using Latent Class Models", Intelligent Systems and Computing, 2020

Conventional data:

- Each **object** is characterised by several **variables**
- Data is organized as an $n \times p$ matrix: rows correspond to **objects**, columns to **variables**

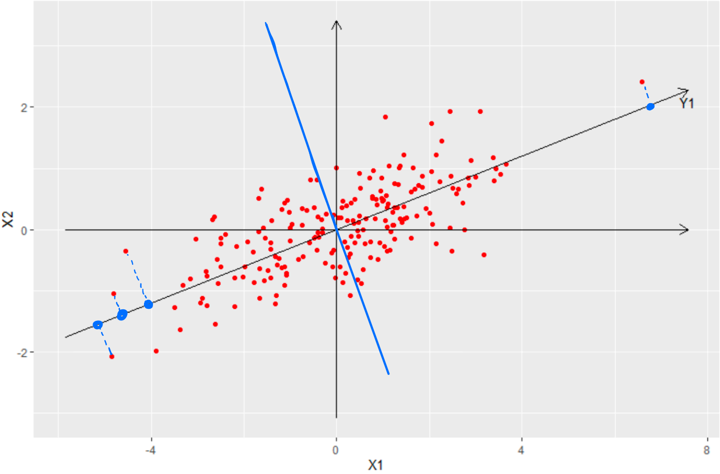
$$\text{Data matrix} \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \qquad \text{random vector } \boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \to \mathbb{R}^p$$

Example: $\begin{bmatrix} x_{11} & \cdots & x_{1p} \end{bmatrix}^T \in \mathbb{R}^p$ is a realisation of $\boldsymbol{X}$.



Y1
X2
2
0
-2
-4
0
4
8
X1

Work with these blue data instead!



Principal Component Analysis (PCA)

covariance matrix $\quad \boldsymbol{\Sigma} = \left[\mathrm{cov}(X_i, X_j)\right] \qquad i, j = 1, \ldots, p$

$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T \qquad \boldsymbol{\Sigma}$ is symmetric

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \qquad \leftarrow$ eigenvalues of $\boldsymbol{\Sigma}$

$\boldsymbol{\Gamma} = [\gamma_1 | \gamma_2 | \ldots | \gamma_p] \qquad \leftarrow$ eigenvectors $\quad \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \boldsymbol{I} = \boldsymbol{\Gamma}^T\boldsymbol{\Gamma}$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \qquad \leftarrow \text{eigenvalues of } \boldsymbol{\Sigma}$$

$$\boldsymbol{\Gamma} = [\gamma_1|\gamma_2|\dots|\gamma_p] \qquad \leftarrow \text{eigenvectors in } \mathbb{R}^p$$

$$\boldsymbol{\Gamma}^T \boldsymbol{\Sigma} \boldsymbol{\Gamma} = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_p) \quad \Rightarrow \quad \gamma_i^T \boldsymbol{\Sigma} \gamma_i = \lambda_i, \quad i = 1, \dots, p.$$

$$\gamma_1 = \arg\max_{\|\gamma\|=1} \gamma^T \boldsymbol{\Sigma} \gamma$$

$$\begin{cases} \gamma_j = \arg\max_{\|\gamma\|=1} \gamma^T \boldsymbol{\Sigma} \gamma \\ \gamma^T \gamma_i = 0, \quad i = 1, \dots, j-1, \end{cases} \qquad j > 1$$

principal components of $\boldsymbol{X} \in \mathbb{R}^p$

$$Y_1 = \gamma_1^T \boldsymbol{X}, \quad Y_2 = \gamma_2^T \boldsymbol{X}, \quad \dots, \quad Y_p = \gamma_p^T \boldsymbol{X}$$

Symbolic data:

How to reduce dimensionality/How to find an SPCA?

- Data is organized as an $n \times p$ matrix, where rows correspond to **objects**, columns to **interval-valued variables**:

$$\text{Data matrix} \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}, \qquad \text{e.g., } x_{11} = [a_{11}, b_{11}]$$

$$\text{random vector } \boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \to \mathbb{IR}^p$$

Example: $\begin{bmatrix} x_{11} & \cdots & x_{1p} \end{bmatrix}^T \in \mathbb{IR}^p$ is a realisation of X.

What do we need?

$\mathbb{IR}^p$ **Principal components of $\boldsymbol{X}$**

$$Y_1 = \gamma_1^T \boldsymbol{X}, \quad Y_2 = \gamma_2^T \boldsymbol{X}, \quad \ldots, \quad Y_p = \gamma_p^T \boldsymbol{X}$$

$$Y_1 = \gamma_1^T \boldsymbol{X} = \gamma_1^T \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} = \gamma_1^T \begin{bmatrix} [a_1, b_1] \\ \vdots \\ [a_p, b_p] \end{bmatrix}$$

1. Define $\mathbb{IR}^p$
2. interval arithmetic, linear combinations of intervals
3. "orthogonality"

- Definition of $\mathbb{IR}^p$

Let

$$\mathbb{IR} = \{[a, b] : a, b \in \mathbb{R}, a \leq b\}$$

be the set of all real closed and bounded intervals,

$[a, b]$ seen as a point (C, R) in $\mathbb{R}^2$

$$C = \frac{a+b}{2} \qquad \text{centre of } [a, b]$$

$$R = b - a \qquad \text{range of } [a, b]$$

- Definition of $\mathbb{IR}^p$

Let

$$\mathbb{IR}^p = \mathbb{IR} \times \mathbb{IR} \times \cdots \times \mathbb{IR}$$

be the cartesian product of p copies of $\mathbb{IR}$.

$$\mathbb{IR} \longleftrightarrow \mathbb{R}^2$$

Let $\mathbf{X} = (X_1, ..., X_p)^T$ be an **interval-valued random vector**

$$X_1 \quad \leftrightarrow \quad (C_1, R_1), \qquad \ldots, \qquad X_p \quad \leftrightarrow \quad (C_p, R_p)$$

$$\boldsymbol{C} = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_p \end{bmatrix} \qquad \boldsymbol{R} = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_p \end{bmatrix}$$

1. Definition of $\mathbb{IR}^p = \mathbb{IR} \times \mathbb{IR} \times \cdots \times \mathbb{IR} \longleftrightarrow \mathbb{R}^{2p}$

Let $\mathbf{X} = (X_1, ..., X_p)^T = (\boldsymbol{C}^T, \boldsymbol{R}^T)^T = \begin{bmatrix} \boldsymbol{C} \\ \boldsymbol{R} \end{bmatrix}$ be an **interval-valued random vector** with realisations in $\mathbb{R}^{2p}$

$$X_1 \quad \leftrightarrow \quad (C_1, R_1), \qquad \ldots, \qquad X_p \quad \leftrightarrow \quad (C_p, R_p)$$

$$\boldsymbol{C} = \begin{bmatrix} C_1 \\ \vdots \\ C_p \end{bmatrix} = \frac{1}{2}\left(\begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix} \right) = \frac{1}{2}(\boldsymbol{A} + \boldsymbol{B}) \qquad \text{random vector of centres}$$

$$\boldsymbol{R} = \begin{bmatrix} R_1 \\ \vdots \\ R_p \end{bmatrix} = \boldsymbol{B} - \boldsymbol{A} \qquad \text{random vector of ranges}$$

- **Moore's interval arithmetic**

For $X = [a, b] = (C, R)$, $\quad Y = [c, d] = (C', R')$ in $\mathbb{IR}$, define

$$X + Y = (C + C', R + R') \qquad = \{x + y \colon x \in X, y \in Y\}$$

$$\alpha X = (\alpha C, |\alpha| R)$$

Linear combination

For $\boldsymbol{X} \in \mathbb{IR}^p$ and $\boldsymbol{\alpha} \in \mathbb{R}^p$, define

$$\boldsymbol{\alpha}^T \boldsymbol{X} = \sum_{i=1}^{p} \alpha_i X_i = (\boldsymbol{\alpha}^T \boldsymbol{C}, |\boldsymbol{\alpha}|^T \boldsymbol{R})^T = \Big(\sum_{i=1}^{p} \alpha_i C_i, \sum_{i=1}^{p} |\alpha_i| R_i\Big)$$

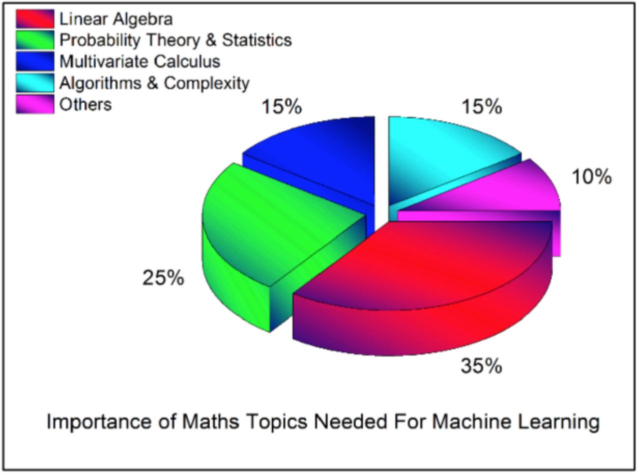
Theorem

The p orthogonal-pcs of $\boldsymbol{X}$ *are* $Y_1, \cdots, Y_p$ *with* $Y_i = \gamma_i^T \boldsymbol{X}$, *where* $\gamma_i \in \mathbb{R}^p$ *is*

$$\gamma_i = \arg\max_{l=1,\ldots,2^p} \left(\gamma_{il}^T \boldsymbol{\Sigma}_{CC} \gamma_{il} + \gamma_{il}^T \boldsymbol{M}_{RR} \gamma_{il} \right) ,$$

and γ_{il}, $l = 1, \ldots, 2^p$, *solves a subproblem:*

$$\gamma_{il} = \begin{cases} \arg\max_{\gamma:||\gamma||=1} \left(\gamma^T \boldsymbol{\Sigma}_{CC} \gamma + \gamma_{il}^T \boldsymbol{M}_{RR} \gamma_{il} \right) \\ \gamma^T \gamma_j = 0, \; j = 1, \ldots, i-1 \;\; \text{if} \;\; i > 1 \end{cases} ,$$



(by Wale Akinfaderin, `https://tinyurl.com/y6hc9sh8`)