Graphical exploratory analysis of educational knowledge surveys with missing and conflictive answers using evolutionary techniques

Luciano Sánchez, José Otero, Inés Couso

Universidad de Oviedo

June 2010

Summary

Introduction: Knowledge Surveys

Initial survey: "System Identification: Regression, Prediction and Time Series"

Name (optional): _____

1. Score yourself (between A and E). How much do you know about the following concepts?
   (a) Control theory:
   ☐ Dynamical system ☐ Linear difference equations ☐ Linear least squares method
   ☐ Impulse response ☐ Transfer function ☐ State-space model
   ☐ Time invariant Kalman filter ☐ Fourier transform ☐ Laplace transform
   (b) Statistics:
   ☐ Random variable ☐ Conditional probability ☐ Conditional expectation
   ☐ Density function ☐ Cumulative distribution function ☐ Radon-Nikodym derivative
   ☐ Likelihood ☐ Shannon Entropy ☐ Regularization
   (c) Numerical Analysis:
   ☐ Gradient ☐ Hessian ☐ Descent algorithm
   ☐ Linear search ☐ Gradient descent ☐ Conjugate gradient
   ☐ Newton's algorithm ☐ Quasi-Newton algorithms ☐ Simulated annealing
   (d) Linear Algebra:
   ☐ Positive definite matrix ☐ Moore-Penrose pseudoinverse ☐ Eigenvector
   ☐ Diagonalization of a matrix ☐ Jacobi Method ☐ Singular Value Decomposition
2. For the practical part of the course, you prefer
   (a) Being taught theory and practice at separate rooms
   (b) Combining theory and practice at the same room, using your own portable computer
   (c) Being taught only theory, using the computer after class by yourself
3. You have chosen this course because
   (a) You want to learn about system identification and modeling
   (b) You want to solve engineering problems of system identification and modeling
4. Comments / suggestions:

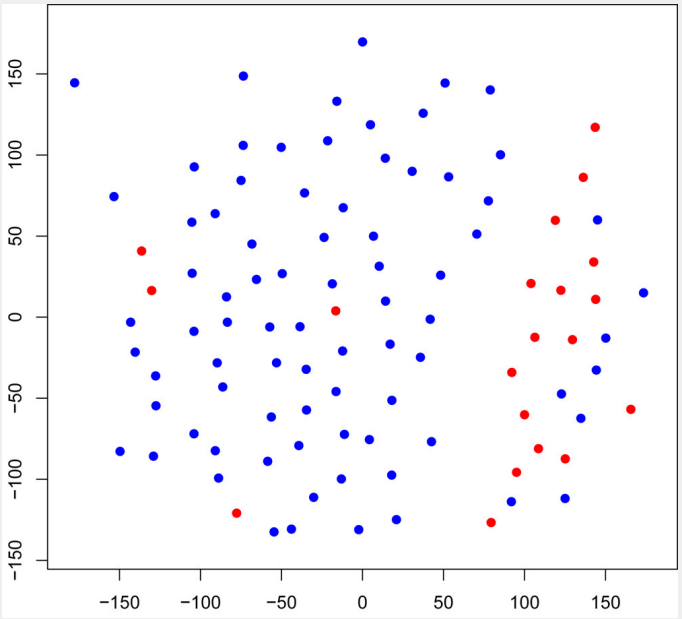
1

- Knowledge surveys comprise short questions that students can answer by writing a single line, or choosing between several alternatives in a printed or web-based questionnaire.
- A survey is not an exam. There is not a high correlation between methodology or dedication and scoring.

Introduction: What KS are for?

- Surveys can be used for deciding the best starting level for the lectures.
- In heterogeneous groups, we can segment the students according to their preparation and establish a common ground.
- When done at the end of the course, the teaching methodology and also the attitude and dedication of the students are assessed.

KSs and graphical analysis

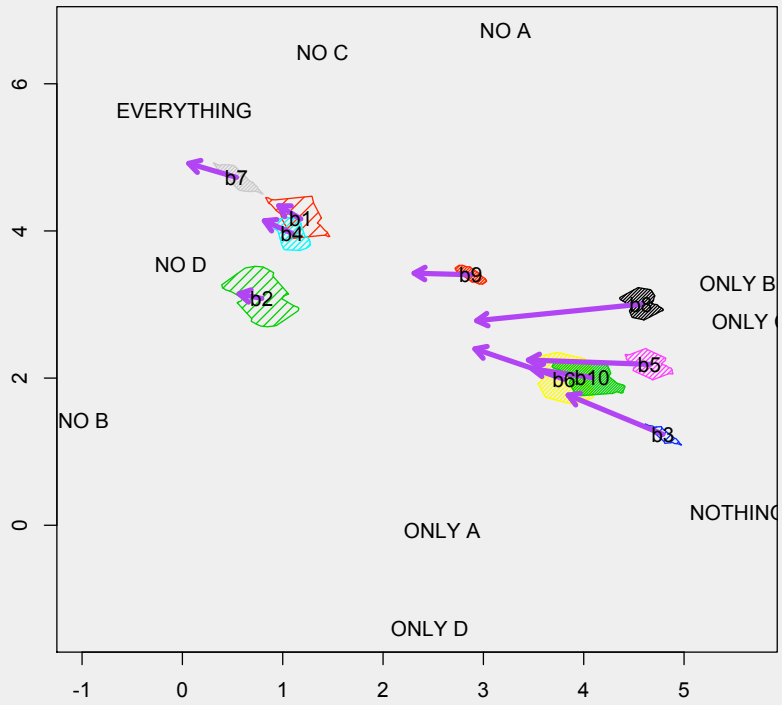


- This paper is about graphically analyzing the data that is collected in a knowledge survey.
- We want to summarize a group of students' learning needs at the beginning of the course and also the capacities acquired during the course.
- Data is projected in a map, where each student will be placed according to his/her knowledge profile
- We have extended this map to data that is incomplete or imprecise.

Conflictive tests and imprecise data

- Why do care about incomplete and imprecise data?
  - The student does not answer some questions of the survey
  - There are incompatible answers that might have been carelessly answered.
- An incomplete survey may be represented by the set of all surveys with any valid value in place of the missing answer.
- For example, if there is a missing answer that should be a number between 0 and 10, the answer to that question will be set to the interval [0,10].
- An incoherent set of answers will also be represented by an interval or a fuzzy set, whose membership function models the dispersion of the conflictive results.

Interpretation of a graphical map

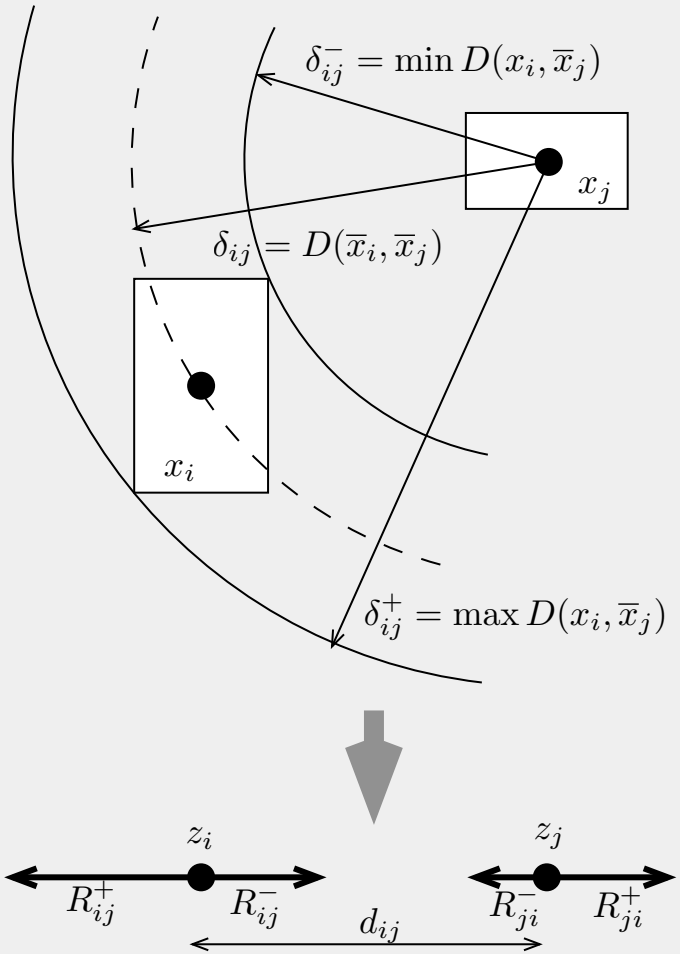


- Each individual in the map is not a point but a figure.
- The relative positions of the figures determine the similarities between it and the other students.
- Their shapes and sizes show the coherence of the answers.
- The displacement of the shapes in time depicts the acquisition of knowledge by the group.

Extension of MDS to fuzzy data

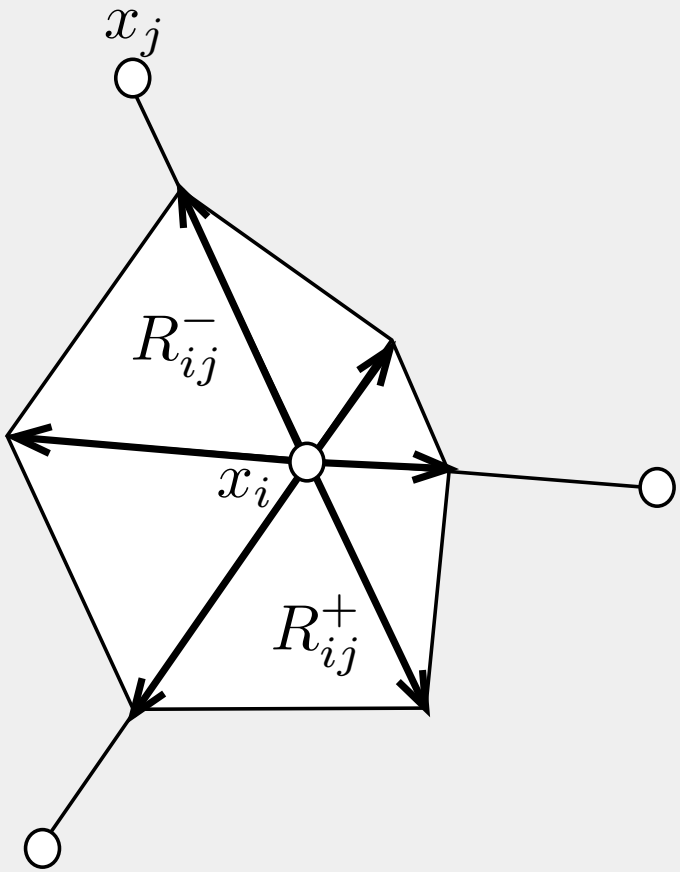
- PCA, SOM, MDS, and other methods project the instances as points in a low dimensional Euclidean space
- MDS consists in finding a two-dimensional cloud of points that minimizes an stress function measuring the difference between the matrix of distances among the data and the matrix of distances of this last cloud.
- The fuzzy extension of this algorithm defines a fuzzy valued stress function that bounds the difference between the imprecisely known matrix of distances among the objects and the fuzzy valued distance matrix between a set of shapes in the low-dimensional projection.
- The coordinates of the elements of the map minimizing the fuzzy stress function will be found with a special purpose GA.

Fuzzy MDS: Distance between cases



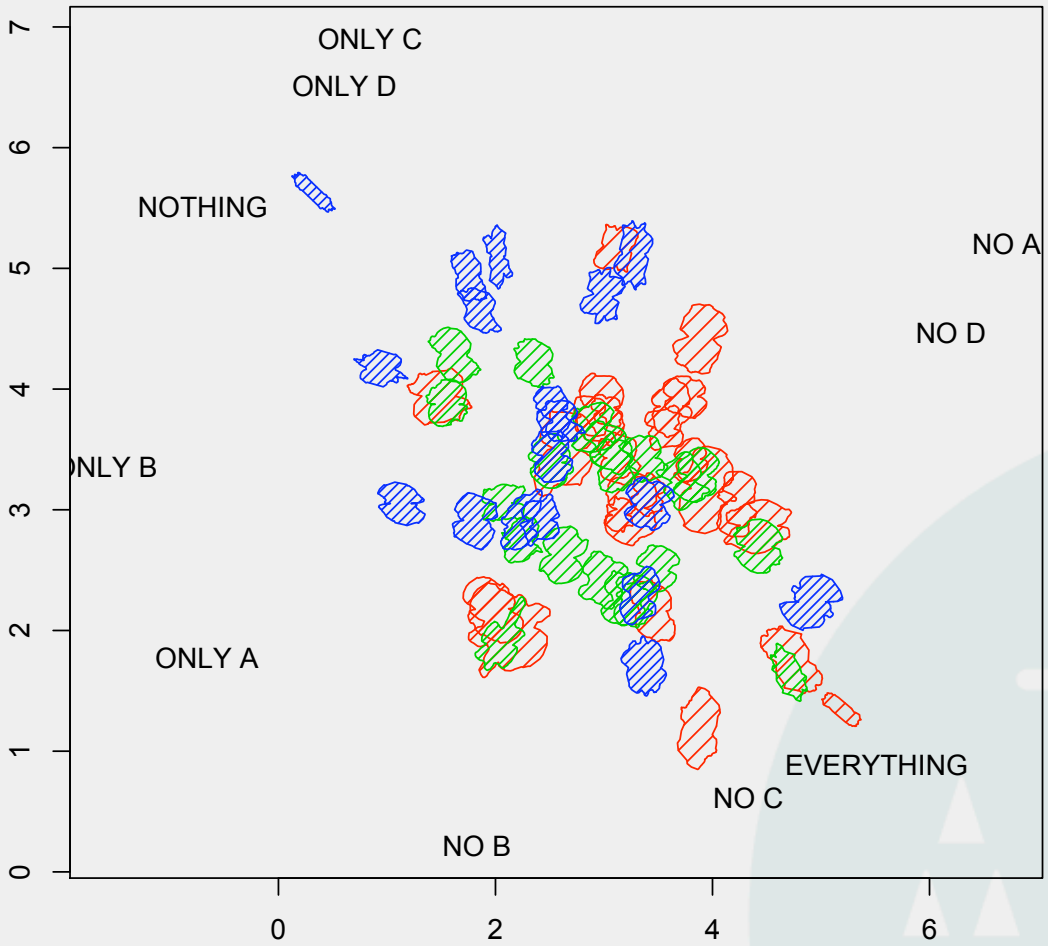
- The distance between two sets is a set itself.
- For instance, the distance between two multivariate interval values ranges between $d_{ij} - R_{ij}^- - R_{ji}^-$ and $d_{ij} + R_{ij}^+ - R_{ji}^+$.

Fuzzy MDS: Shapes of the projection

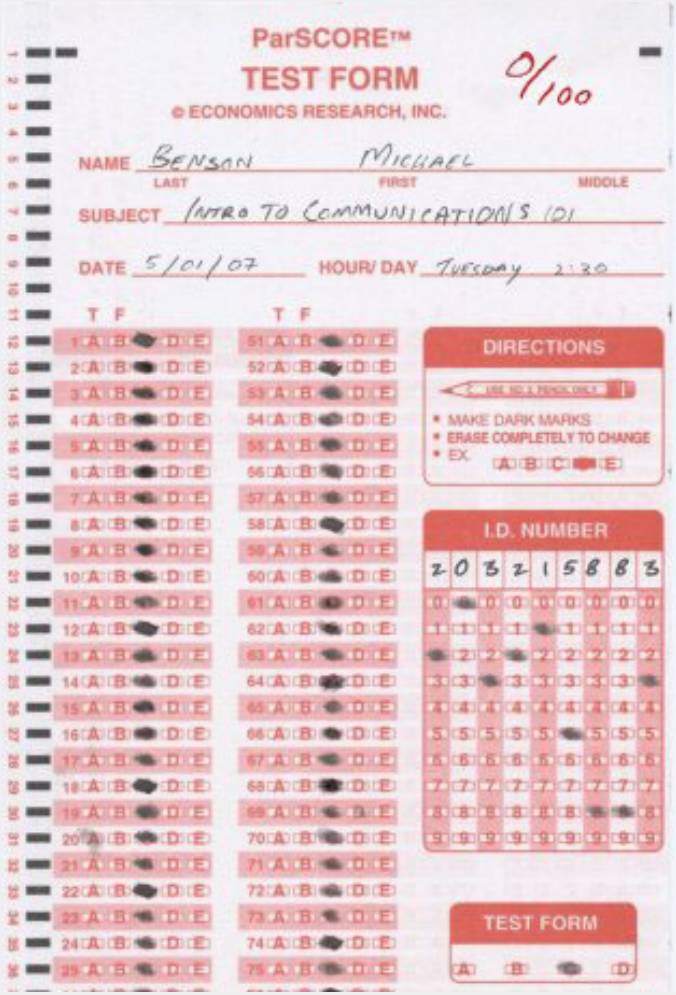


- The approximation to the shape of an object is a polygon whose vertexes are in the lines joining the centers of the projections.
- The value of the stress function our map has to minimize is $\sum_{i=1}^{N}\sum_{j=i+1}^{N} d_H(D_{ij}, [d_{ij} - R_{ij}^- - R_{ji}^-, d_{ij} + R_{ij}^+ + R_{ji}]^+)^2$

Fuzzy MDS: Shapes of the projection

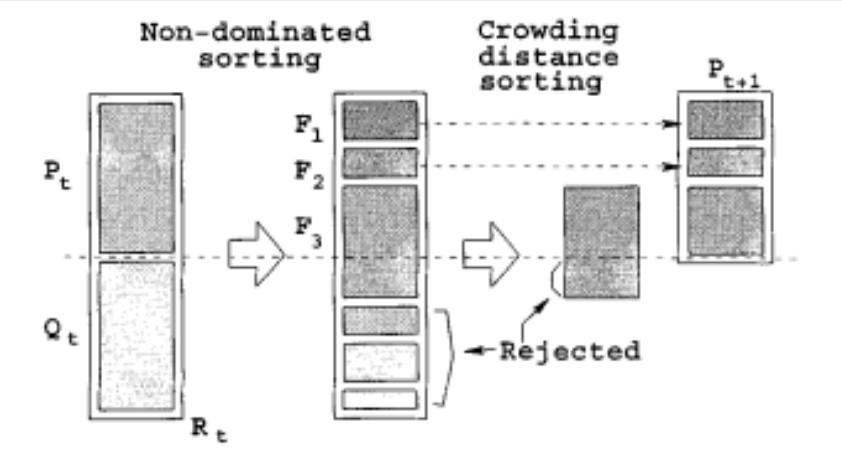


Characteristic points



- We make up prototypes of students (no mistakes, a completely wrong survey, just one correct answer, etc.)
- The capacities of a student can be related to those of his closest characteristic point.

Evolutionary algorithm



- An evolutionary algorithm is used for optimizing the stress function and searching the best map.
- Interval and fuzzy fitness functions can be optimized with extensions of multiobjective genetic algorithms. In this paper we have used the extended NGSA-II.

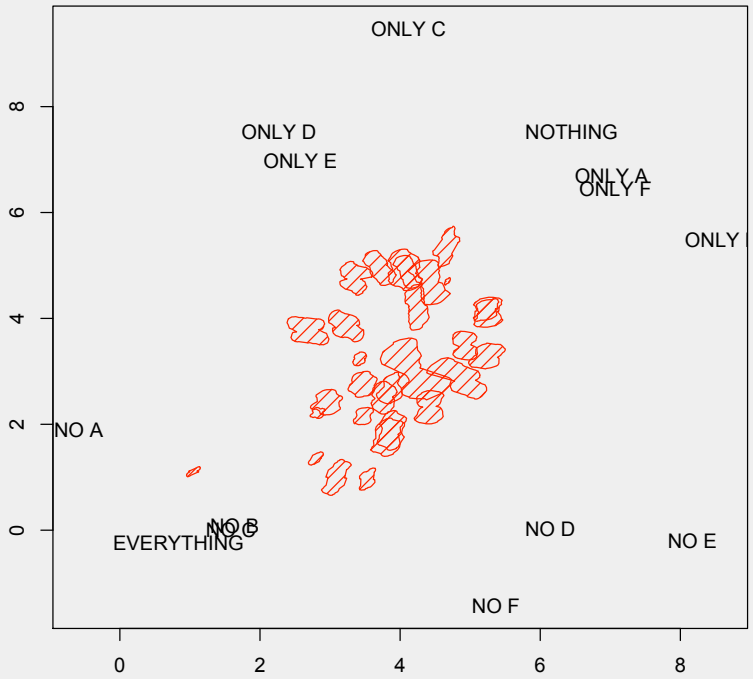
Coding Scheme

- Each individual of the population represents a set of coordinates in the plane.
- Each chromosome consists of the concatenation of so many pairs of numbers as students, plus one pair for each characteristic point (i.e. "Everything", "Nothing", "Only Subject X", "Every Subject but X", etc).
- The chromosome is fixed-length, and real coding is used.
- Arithmetic crossover is used for combining two chains. The mutation operator consists in performing crossover with a randomly generated chain.

Evolutionary scheme for a fuzzy fitness function

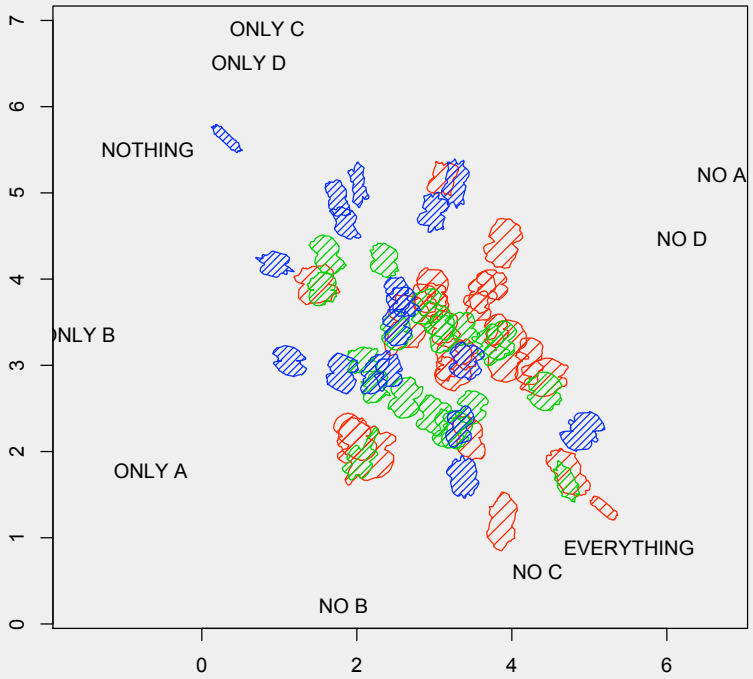
- Generational approach with the multiobjective NSGA-II replacement strategy.
- Binary tournament selection based on the crowding distance in the objective function.
- Precedence operator derived from the bayesian coherent inference with an imprecise prior.
- Non-dominated sorting based on the product of the lower probabilities of precedence.
- Crowding based on the Hausdorff distance.

Real-world examples (I)



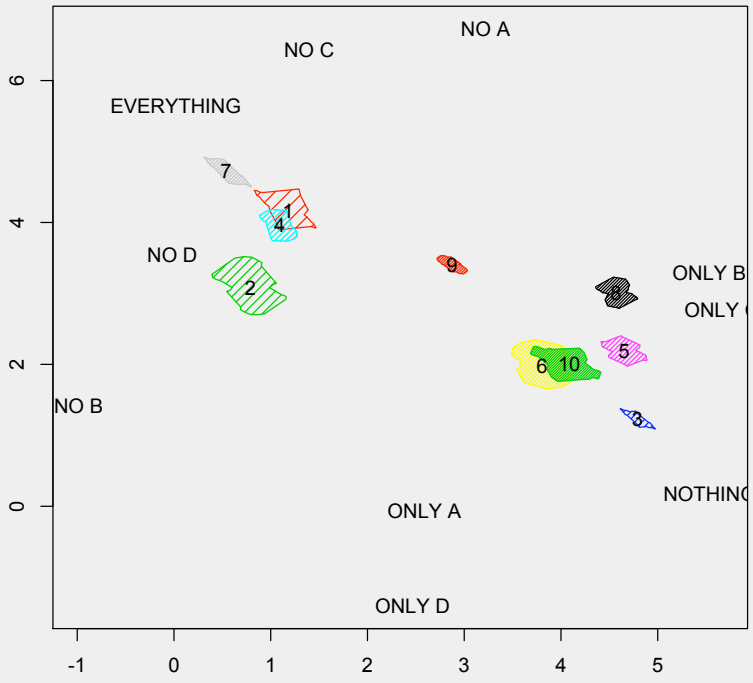
- 30 students of subject "Statistics', beginning of 2009-2010 course.
- This survey evaluates previous knowledge in Algebra (A), Logic (B), Electronics (C), Numerical Analysis (D), Probability (E) and Physics (F).
- The positions of the characteristic points have been marked with labels. "A" means that all the questions about the subject "A" are correct "NO A" means that all the questions except "A" ones are correct, etc.

Real-world examples (II)



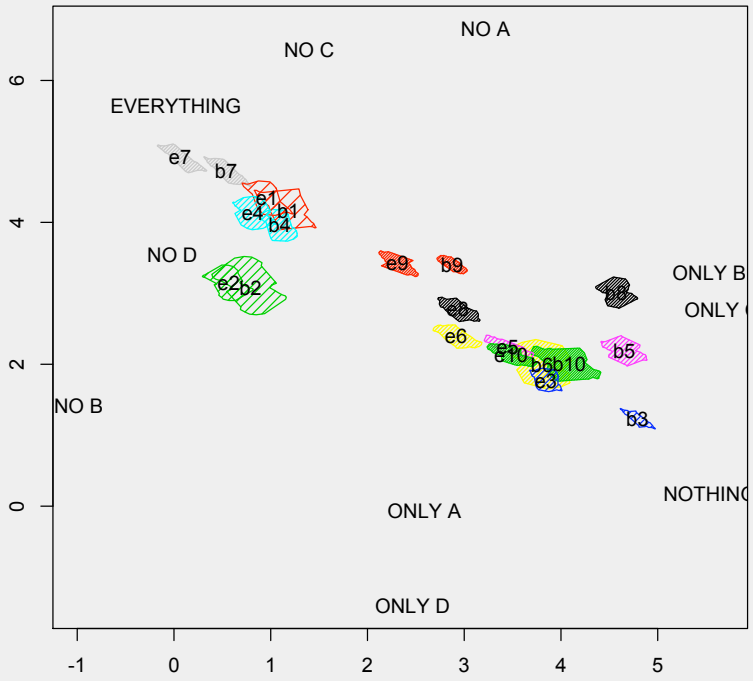
- Knowledge about prerrequisites of Computer Science, engineering students specialized in Chemistry (green), Electricity (red) and Mechanics (blue).
- The students of the intensification coded in red consider themselves better prepared than those coded in blue, with the green group in an intermediate position, closer to red.
- All the students of all the groups have a neutral orientation to math subjects.

Real-world examples (III)



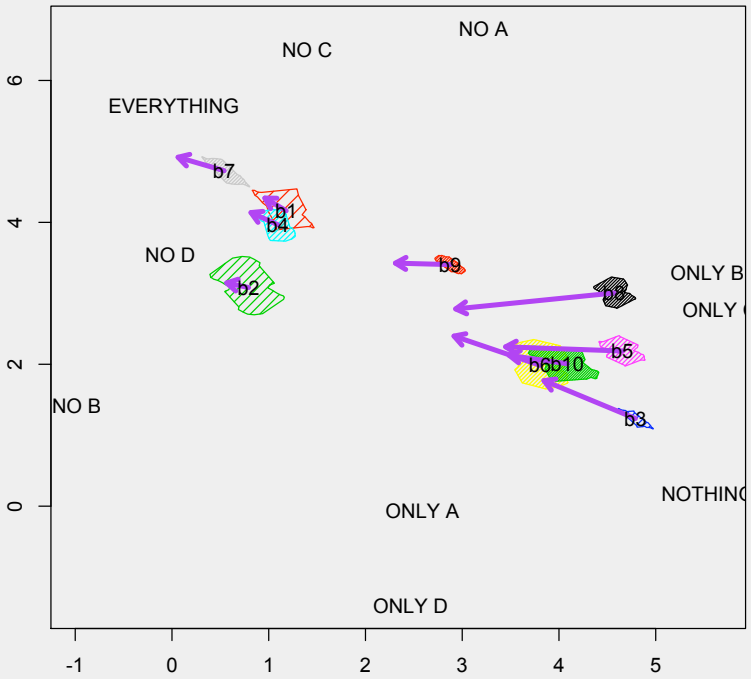
- Research master. Ten pre-doctoral students in Computer Science, Physics and Mathematics.
- 36 subjects classified in "Control Algorithms" (A), "Statistical Data Analysis" (B), "Numerical Algorithms" (C) and "Lineal Models" (D). Students from technical degrees (Computer Science) evaluated themselves with the lowest scores (shapes in the right part of each figure).

Real-world examples (IV)



- The same survey, end of the course.
- All the students moved to the left, closer to characteristic point "EVERYTHING".
- The displacement has been larger for the students in the group at the right.

Real-world examples (V)



- Shapes obtained from the final survey replaced by arrows that begin in the initial position and end in the final center.
- The length of the arrows is related with the progress of the student during the course.

Concluding remarks

- Extension of Multidimensional Scaling to imprecise data, applied to a map that summarizes incomplete or carelessly filled surveys that may include conflictive answers.
- The map of a group of students consists on several shapes, whose volume measures the degree to which a survey lacks consistency.
- These maps can help detecting heterogeneous groups and can also be used for assessing the results of a course.