Dimensionality reduction and visualization of fuzzy data

A Survey

Thierry Denœux[1] and Marie-Hélène Masson[1]

[1]Université de Technologie de Compiègne
HEUDIASYC (UMR CNRS 6599)

ISI '07 - IPM30 Interval and Imprecise Data

Motivations
Objectives

- One of the main tasks in exploratory data analysis: search for a relevant low-dimensional feature space in which the original data can be mapped and displayed so as to uncover their underlying structure.
- Usually, the data are precise (each observation consists in a single value), and each object is represented as a point of $\mathbb{R}^q$ (precise representation).
- Problem: how to extend in a meaningful way the usual feature extraction and data visualization methods to handle imprecise data?
- Reasonable requirement: when data are imprecise, each object should have an imprecise representation as a region of $\mathbb{R}^q$.

Motivations
Imprecise data

- By "imprecise data", we mean set-valued observations, ie, observations consisting in crisp or fuzzy sets of values.
- Two main cases for numerical data:
  - Interval-valued data: data items are (crisp) real intervals;
  - Fuzzy data: data items are fuzzy intervals (roughly, real intervals with ill-defined bounds).
- Each crisp or fuzzy set of values represents
  - either an imprecise (partial) observation of some precise unknown quantity (e.g., temperature in this room is "around 20 °C", or in the range $[19, 21]$), or
  - a distribution of values obtained from repeated measurements, or related to different entities forming a class of interest.

Motivations

Object-attribute vs. dissimilarity data

- Feature extraction can be divided in two subproblems:
  - feature extraction from object-attribute data: transform an $n \times p$ data matrix $\mathbf{X}$ whose rows are p-dimensional feature vectors observed for n objects, into a matrix $\mathbf{Y}$ of size $n \times q$, with $q < p$. Classical approach: principal component analysis (PCA).
  - feature extraction from dissimilarity data: given an $n \times n$ matrix $\Delta = (\delta_{ij})$ of pairwise dissimilarities between n objects, finds a $n \times q$ data matrix $\mathbf{Y}$ of n points in a q dimensional space such that the interpoint distances reflect the input dissimilarities. Classical approach: multidimensional scaling (MDS).
- We will focus on the extension of PCA and MDS to imprecise object-attribute and dissimilarity data, respectively.

Overview

Overview

PCA

Principles and notations (1/2)

- Let $\mathbf{X} = (x_{ij})$ be the numerical data matrix of order $(n \times p)$, where n denotes the number of objects, and p the number of variables.
- We assume $\mathbf{X}$ to be centered, i.e., $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$ for all $j \in \{1, \ldots, p\}$.
- We can think of the n data points as a cloud in $\mathbb{R}^p$, with center of gravity located at the origin.
- PCA attempts to find a q-dimensional subspace $\mathcal{L}$ of $\mathbb{R}^p$, with $q \leq p$, such that the orthogonal projections of the n points on $\mathcal{L}$ have maximal variance.

PCA

Principles and notations (2/2)

- The solution is known to be the subspace spanned by the q normalized eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_q$ of the sample covariance matrix $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}$, associated with the first q largest eigenvalues.
- The matrix $\mathbf{U}_q = (\mathbf{u}_1, \ldots, \mathbf{u}_q)$ of order $(p \times q)$ is sometimes called the component loading matrix.
- The coordinates of the objects in the projected space are defined by matrix $\mathbf{Y} = \mathbf{X}\mathbf{U}_q$, often referred to as the component score matrix.

Overview

Centers PCA

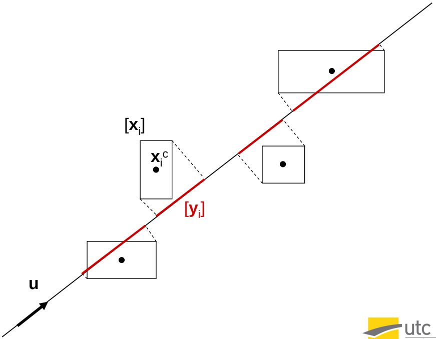
Extension to interval data

- Let us now assume that we have an interval data matrix $[\mathbf{X}] = ([x_{ij}^-, x_{ij}^+])$ of size $n \times p$.
- Each line of $[\mathbf{X}]$ is a vector $[\mathbf{x}_i] = ([x_{i1}], \ldots, [x_{ip}])$ of intervals called a box. It may be identified with the region of $\mathbb{R}^p$ defined by $[x_{i1}] \times \ldots \times [x_{ip}]$.
- The simplest extension of PCA to such data was introduced by Cazes et al (1997): Centers PCA (C-PCA).
- Basic idea: apply standard PCA to the single-valued data matrix $\mathbf{X}^c$ obtained by replacing each interval $[x_{ij}^-, x_{ij}^+]$ by its center $x_{ij}^c = (x_{ij}^- + x_{ij}^+)/2$.
- Let $\mathbf{U}_q = (\mathbf{u}_1, \ldots, \mathbf{u}_q)$ be the corresponding component loading matrix.

Centers PCA

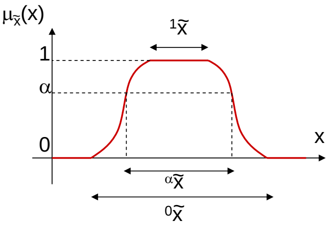
Definition of interval-valued component scores

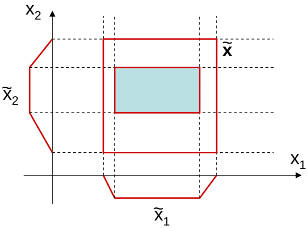
- We may define the interval-valued component scores for object i as the bounds of the component scores for all $\mathbf{x} \in [\mathbf{x}_i]$.
- Each box $[\mathbf{y}_i]$ is the interval hull of the set of component scores of the vertices of $[\mathbf{x}_i]$.
- Matrix $[\mathbf{Y}]$ is easily computed as $[\mathbf{Y}] = [\mathbf{X}]\mathbf{U}_q$ using interval arithmetics.



Centers PCA
Extension to fuzzy data

- Let $\widetilde{\mathbf{X}} = (\widetilde{x}_{ij})$ be a fuzzy data matrix of size $n \times p$.
- Each data item $\widetilde{x}_{ij}$ is a fuzzy interval, ie, a fuzzy subset of $\mathbb{R}$ whose α-cuts ${}^{\alpha}\widetilde{x}_{ij}$ for $\alpha \in (0, 1]$ are closed intervals.
- Each line of $\widetilde{\mathbf{X}}$ is a vector $\widetilde{\mathbf{x}}_i = (\widetilde{x}_{i1}, \ldots, \widetilde{x}_{ip})$ of fuzzy intervals that may be called a fuzzy box. It can be identified with the fuzzy subset of $\mathbb{R}^p$ with α-cuts ${}^{\alpha}\widetilde{\mathbf{x}}_i = {}^{\alpha}\widetilde{x}_{i1} \times \ldots {}^{\alpha}\widetilde{x}_{ip}$.

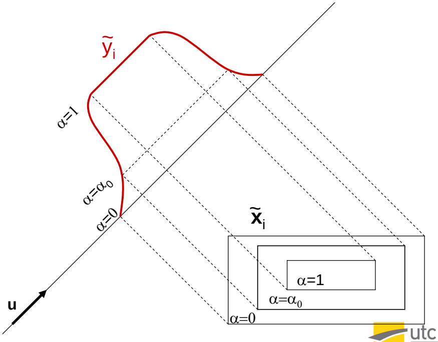




Centers PCA

Extension to (fuzzy data)

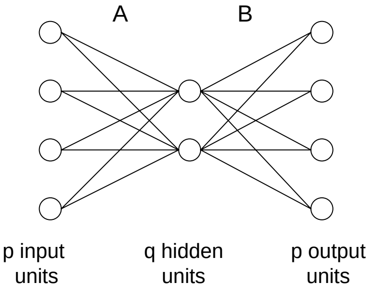
- Let $\mathbf{U}_q$ be the component loading matrix obtained by applying standard PCA to the defuzzified data matrix;
- Fuzzy component scores may be defined by projecting each α-cut ${}^{\alpha}\widetilde{\mathbf{x}}_i$ on $\mathcal{L}$.
- If all fuzzy numbers are trapezoidal, matrix $\widetilde{\mathbf{Y}}$ is easily computed as $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{X}}\mathbf{U}_q$ using fuzzy arithmetics.



Overview

Connection between PCA and neural networks
NN architecture

- As first noticed by Bourlard and Kamp (1988), there is an interesting connection between PCA and autoassociative multilayer perceptrons (MLPs).
- Let us consider a three-layer MLP:



Connection between PCA and neural networks

Error function

- Let us assume that this network is trained in autoassociative mode, i.e., using the inputs as target outputs, with the quadratic error function:

$$E(A,B) = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{z}_i\|^2 = \sum_{k=1}^{p} (x_{ik} - z_{ik})^2,$$

  where $\mathbf{z}_i = BA\mathbf{x}_i$ is the vector of outputs for input vector $\mathbf{x}_i$.
- As $E(A,B) = E(CA, BC^{-1})$ for any invertible $q \times q$ matrix C, the error may be expressed as a function of the global map $W = BA$.

Connection between PCA and neural networks

Main result

Theorem (Baldi and Hornik, 1989)

The error E expressed as a function of the global map W has a unique local and global minimum of the form $W = BA$ with

$$
\begin{aligned}
A &= CU'_q \\
B &= U_q C^{-1},
\end{aligned}
$$

where U_q is the component loading matrix and C is an arbitrary invertible $q \times q$ matrix.

Connection between PCA and neural networks

Hidden unit activities as transformed component scores

- The vector of hidden unit activities is then

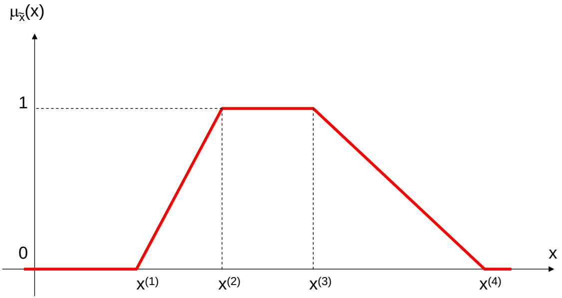
$$A\mathbf{x} = CU'_q\mathbf{x} = C\mathbf{y}$$

  where $\mathbf{y}$ is the vector of component scores for input $\mathbf{x}$ $\rightarrow$ it is identical to $\mathbf{y}$ up to an arbitrary linear transformation.
- If the constraint $A' = B$ is imposed, then C is becomes an orthogonal matrix: the hidden unit activities and the principal components are then related by an isometric transformation.
- The propagation equation becomes $\mathbf{z} = BB'\mathbf{x}$.

NN-PCA

Extension to fuzzy input data

- Let us now assume that we have a fuzzy data matrix $\widetilde{\mathbf{X}} = (\widetilde{x}_{ij})$.
- Each data item $\widetilde{x}_{ij}$ will be assumed to be a trapezoidal fuzzy number parameterized as $\widetilde{x}_{ij} = (x_{ij}^{(1)}, x_{ij}^{(2)}, x_{ij}^{(3)}, x_{ij}^{(4)})$:



NN-PCA

Propagation of fuzzy inputs

- Input data can be propagated in the PCA autoassociative neural network using Zadeh's extension principle (a principle for extending any function to fuzzy sets).
- The calculations can be easily performed using fuzzy arithmetics.
- The vectors $\widetilde{\mathbf{y}}$ and $\widetilde{\mathbf{z}}$ of hidden unit activations and outputs are defined as

$$\begin{aligned} \widetilde{\mathbf{y}} &= B'\widetilde{\mathbf{x}} \\ \widetilde{\mathbf{z}} &= B\widetilde{\mathbf{y}}. \end{aligned}$$

- Their components are trapezoidal fuzzy numbers.

NN-PCA
Error function

- The reconstruction error for component k of input vector i may be defined as

$$e(\widetilde{x}_{ik}, \widetilde{z}_{ik}) = \sum_{\ell=1}^{4} (z_{ik}^{(\ell)} - x_{ik}^{(\ell)})^2, \quad k = 1, \ldots, d, \qquad (1)$$

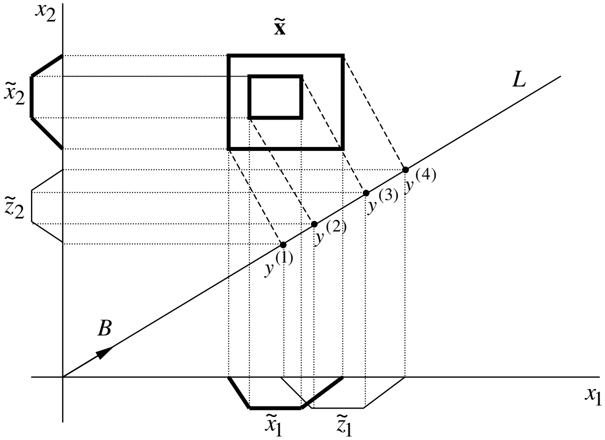
- The total error is

$$E(B) = \sum_{i=1}^{n} \sum_{k=1}^{p} e(\widetilde{x}_{ik}, \widetilde{z}_{ik}).$$

- The minimization of E with respect to B can be performed using a gradient descent procedure.

NN-PCA

Illustration



Overview

Sensory evaluation example

Data and problem statement

- Data from a research project performed in collaboration with a French car manufacturer.
- The entities under study were noises recorded inside several vehicles.
- The data consisted in scores given by 12 judges describing their perception of 21 sounds according to 5 attributes. Each sound was presented 3 times to each subject, yielding a four-way data matrix: sounds $\times$ attributes $\times$ subjects $\times$ replications.
- The aim of this work was to study the variability of the responses among the panelists and the variability of each subject accross repetitions.

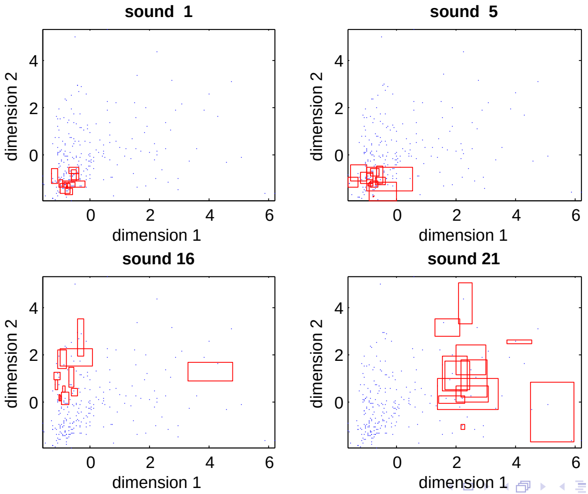
Sensory evaluation example

Data encoding and experimental setting

- Each of the 21×12 pairs (sound, subject) was considered as an object described by five fuzzy attributes.
- For each attribute, the three scores available from replications were converted into a triangular fuzzy number (which is a special case of trapezoidal fuzzy number with $x^{(2)} = x^{(3)}$) defined by the minimum, maximum and median value.
- We thus obtained a set of 12×21 vectors composed of 5 triangular fuzzy numbers.
- An autoassociative network with two hidden units ($q = 2$) was used to visualize the data.

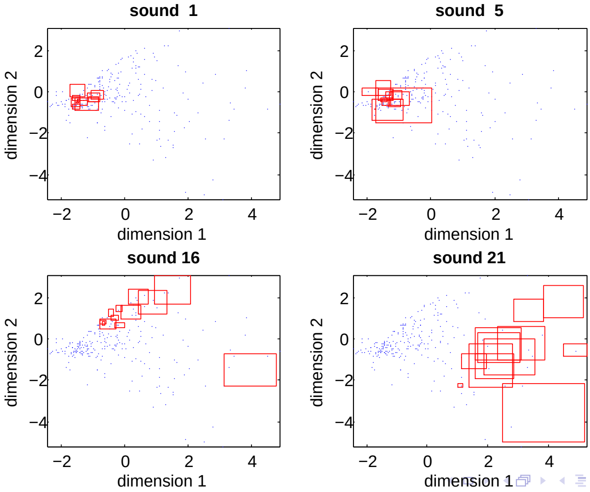
Sensory evaluation example

Two-dimensional projection of sounds (NN-PCA)



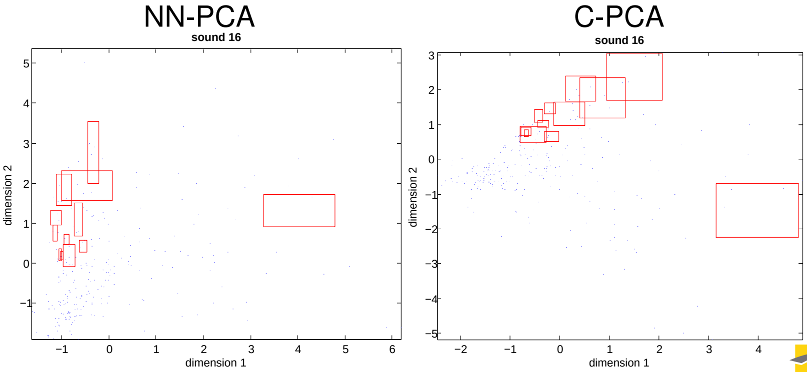
Sensory evaluation example

Two-dimensional projection of sounds (C-PCA)

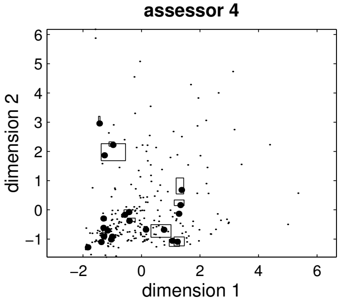


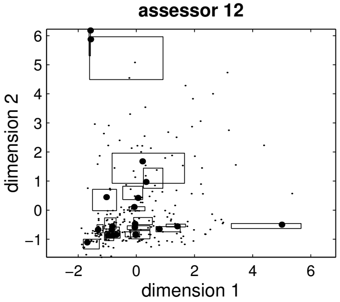
Sensory evaluation example

Comparison between NN-PCA and C-PCA

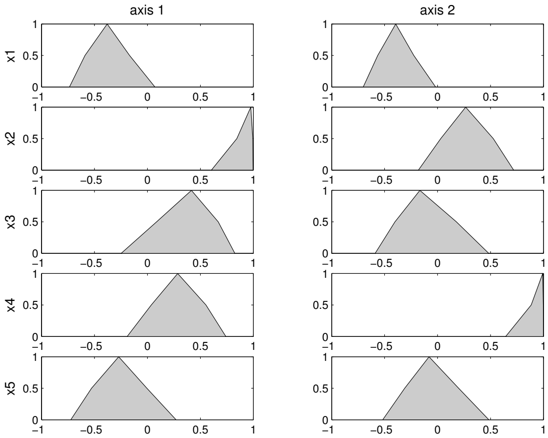


Sensory evaluation example
Comparison of two assessors (NN-PCA)





Sensory evaluation example
Fuzzy correlations



Overview

Multidimensional scaling

Principles and notations

- Let $\Delta = (\delta_{ij})$ be a square matrix expressing the precise dissimilarities between n objects.
- Classically, we seek to represent each object i by a point $\mathbf{x}_i$ in $\mathbb{R}^p$ such that the interpoint distances reflect, according to some criterion, the input dissimilarities.
- Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ be the $n \times p$ matrix encoding the n p-dimensional vectors. We search $\mathbf{X}$ so as to minimize a cost (stress) function such as:

$$\sigma(\mathbf{X}) = \sum_{i<j} (d_{ij} - \delta_{ij})^2,$$

where d_{ij} is the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$.

- $\sigma(\mathbf{X})$ can be minimized using an iterative procedure.

Multidimensional scaling

Generalization to interval data

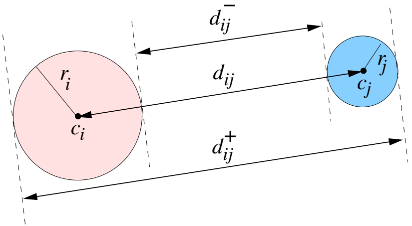
- Let us now assume the dissimilarities to be given in the form of intervals $[\delta_{ij}] = [\delta_{ij}^-, \delta_{ij}^+]$.
- Each interval may be interpreted as the set of possible values for the true unknown dissimilarity δ_{ij}.
- Since the objects are imprecisely located with respect to each other, it is natural to represent object as a regions R_i in $\mathbb{R}^p$.
- The minimum and maximum distances between two regions R_i and R_j are then defined by:

$$
\begin{aligned}
d_{ij}^- &= \min_{\mathbf{x}_i \in R_i, \mathbf{x}_j \in R_j} \|\mathbf{x}_i - \mathbf{x}_j\| \\
d_{ij}^+ &= \max_{\mathbf{x}_i \in R_i, \mathbf{x}_j \in R_j} \|\mathbf{x}_i - \mathbf{x}_j\|.
\end{aligned}
$$

Multidimensional scaling

Hypersphere model

- In the simplest model, each region R_i is chosen to be a hypersphere with center $\mathbf{c}_i \in \mathbb{R}^p$ and radius $r_i \in \mathbb{R}_+$.
- d_{ij}^- and d_{ij}^+ can then be simply obtained as functions of the radii, and the distance d_{ij} between the two centers:



- The problem is then to determine the centers and the radii such that the interval-valued distances represent the dissimilarities in an optimal way.

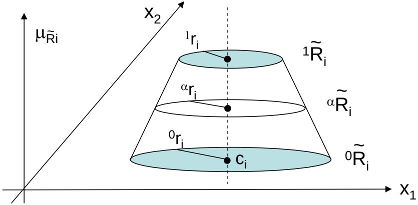
Multidimensional scaling
Generalization to fuzzy data

- More general situation: each dissimilarity is expressed as a fuzzy interval.
- Such data may come from a linguistic evaluation ("very close", "quite different", etc.), or from a distribution of responses from a panel of assessors.
- It is then natural to represent each object by a fuzzy region $\widetilde{R}_i$ in $\mathbb{R}^p$ defined by a fuzzy membership function $\mu_{\widetilde{R}_i}$.
- According to Zadeh's extension principle, the distance between two fuzzy regions $\widetilde{R}_i$ et $\widetilde{R}_j$ can be defined as a fuzzy interval $\widetilde{d}_{ij}$.

Multidimensional scaling

Fuzzy hypersphere model

- Simple model: each object is represented by a fuzzy region whose α-cuts are concentric hyperspheres of radii ${}^{\alpha}r_i$ and center $\mathbf{c}_i$:



- Each α-cut of $\widetilde{d}_{ij}$ is a closed interval ${}^{\alpha}\widetilde{d}_{ij} = [{}^{\alpha}\widetilde{d}_{ij}^{-}, {}^{\alpha}\widetilde{d}_{ij}^{+}]$, whose bounds are the minimum and maximum distances between the α-cuts of $\widetilde{R}_i$ and $\widetilde{R}_j$.

Overview

Least-squares fitting
Interval-valued dissimilarities

- In the case of interval-valued dissimilarities, the stress function can be defined as:

$$\sigma'(\mathcal{R}) = \sum_{i<j}(d_{ij}^- - \delta_{ij}^-)^2 + \sum_{i<j}(d_{ij}^+ - \delta_{ij}^+)^2,$$

  where $\mathcal{R}$ denotes the set of n regions $\{R_1, \ldots, R_n\}$.
- The $n(p+1)$ model parameters (n centers defined by p coordinates and n radii) can then be determined by minimizing $\sigma'(\mathcal{R})$ with respect to $\mathcal{R}$, using an iterative gradient descent algorithm.

Least squares fitting

Properties

It may be shown that:

- If all the dissimilarities are precise (i.e. $\delta_{ij}^- = \delta_{ij}^+$), the model leads to null radii, thereby generalizing the classical model;
- Otherwise, each radius r_k is linearly related to the quantity

$$s_k = \sum_{i \neq k} (\delta_{ik}^+ - \delta_{ik}^-),$$

which is a measure of the global imprecision of the dissimilarities between object k and all other objects.

Consequently, the size of the region R_i describing object i is related to the imprecision of the data regarding that object.

Least-squares fitting
Fuzzy dissimilarities

- To fit the fuzzy model, a set $\{\alpha_i\}_{i=1,c}$ of predetermined levels of α-cuts has to be chosen.
- The stress function can then be defined as:

$$\sigma''(\widetilde{\mathcal{R}}) = \sum_{k=1}^{c} \sum_{i<j} ({}^{\alpha_k}\widetilde{d}_{ij}^{-} - {}^{\alpha_k}\widetilde{\delta}_{ij}^{-})^2 + \sum_{k=1}^{c} \sum_{i<j} ({}^{\alpha_k}\widetilde{d}_{ij}^{+} - {}^{\alpha_k}\widetilde{\delta}_{ij}^{+})^2,$$

  where $\widetilde{\mathcal{R}}$ denotes the set of the fuzzy regions $\widetilde{R}_i$, and ${}^{0}\widetilde{x}$ represents, by convention, the support of fuzzy number $\widetilde{x}$.
- The number of parameters of the model is $n(p+c)$: n centers defined by p coordinates c_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$ and nc radii ${}^{\alpha_k}r_i$, $i = 1, \ldots, n$, $k = 1, \ldots, c$.

Overview

Possibilistic fitting

Interval-valued dissimilarities

- As we have seen, the LS approach provides a configuration such that the dissimilarities are recovered approximatively.
- In contrast, the possibilistic approach searches for a configuration that provides guaranteed bounds for dissimilarities.
- Let us suppose that the centers $\mathbf{c}_i$ have already been determined, e.g. using the LS method.
- We may attempt to find the smallest radii r_i such that the following condition is satisfied:

$$[\delta_{ij}^-, \delta_{ij}^+] \subseteq [d_{ij}^-, d_{ij}^+] \quad \forall i, j.$$

Possibilistic fitting

Formalization as a LP problem

- This leads to the following optimization problem:

$$\min_{\mathbf{r}} \sum_{i=1}^{n} r_i$$

subject to:

$$[\delta_{ij}^-, \delta_{ij}^+] \subseteq [d_{ij}^-, d_{ij}^+] \quad \forall i, j. \tag{2}$$

$$r_i \geq 0 \quad \forall i = 1, n\,. \tag{3}$$

- Constraints (2) may be written

$$r_i + r_j \geq \max(d_{ij} - \delta_{ij}^-, \delta_{ij}^+ - d_{ij}) \quad \forall i, j.$$

- This is a linear programming problem, which is always feasible.

Possibilistic fitting

Interpretation

Remark

In contrast to least squares fitting, possibilistic fitting does not lead to null radii in case of precise but erroneous input dissimilarities: the obtained representation reflects both

- the imprecision in the data (the widths of the input dissimilarities) and
- the goodness-of-fit of the model (i.e., the choice of the Euclidean distance, the dimensionality of the configuration, and the estimation errors).

Possibilistic fitting

Fuzzy dissimilarities

- We now seek fuzzy regions such that $\widetilde{\delta}_{ij} \subseteq \widetilde{d}_{ij}$, $\forall i, j$, where $\subseteq$ now denotes the standard fuzzy set inclusion, i.e.

$$\mu_{\widetilde{\delta}_{ij}} \leq \mu_{\widetilde{d}_{ij}}, \quad \forall i, j.$$

- Again, this may be achieved by a LP problem:

$$\min_{\mathbf{r}} \sum_{k=1}^{c} \sum_{i=1}^{n} {}^{\alpha_k} r_i$$

subject to:

$$^{\alpha_k} r_i + {}^{\alpha_k} r_j \geq \max(d_{ij} - {}^{\alpha_k}\delta_{ij}^{-}, {}^{\alpha_k}\delta_{ij}^{+} - d_{ij}) \quad \forall i, j, k$$

$$^{\alpha_0} r_i \geq 0, \quad \forall i$$

$$^{\alpha_k} r_i \leq {}^{\alpha_{k+1}} r_i, \quad \forall i, \forall k < c,$$

Overview

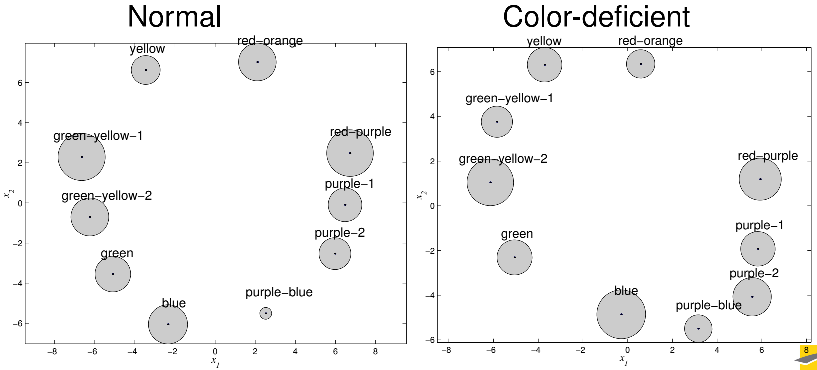
Color data example

Data and experimental settings

- Experiment reported by Helm (1964) about the perception of colors by human subjects.
- Ten colored objects were presented to different subjects who were asked to rate the perceived dissimilarities.
- They were classified into two groups: some of them had a normal color vision, whereas the other had a color-deficient vision. Two separate analyses were conducted on these two groups.
- The perception of each group was summarized using a triangular fuzzy number computed from the minimum, maximum and mean responses of the subjects.

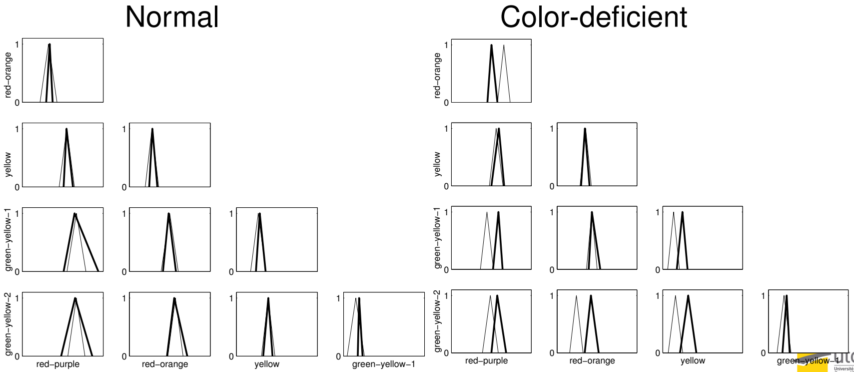
Color data example

Results using the LS model



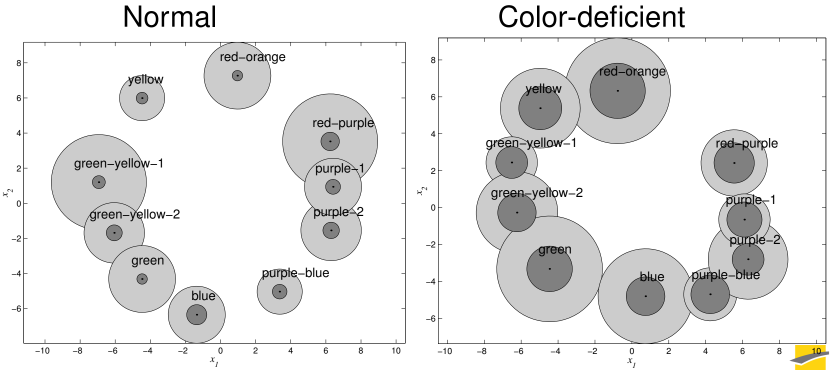
Color data example

Reconstruction of dissimilarities using the LS model



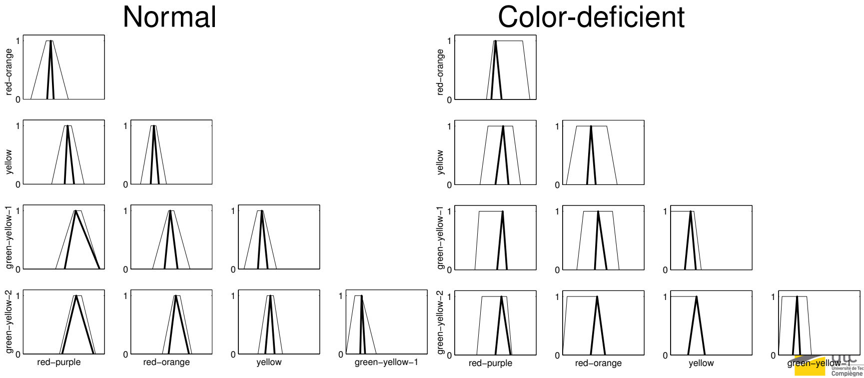
Color data example

Results using the possibilistic model



Color data example

Reconstruction of dissimilarities using the possibilistic model



Conclusions
Summary

- Methods for dimensionality reduction and visualization of interval and fuzzy data have been reviewed.
- These methods extend PCA and MDS in such a way that each object is no longer represented by a point, but by a crisp or a fuzzy region in a low-dimensional feature space.
- This makes it possible to represent in the same display both the variatiability accross objects, but also the imprecision or spread of observations for each object.

Conclusions

Main principles

Three main approaches for extending classical feature extraction and data visualizing techniques to imprecise data:

- Propagation approach: construct a mapping from the original feature space to a lower dimensional feature space using precise data, and then compute the images of interval-valued or fuzzy data through this mapping (C-PCA).
- Cost minimization approach: extend the cost functions optimized by standard approaches, to the case of interval-valued or fuzzy data (NN-PCA, LS-MDS).
- Imprecision minimization approach: find the most precise representations of the original data, verifying some constraints (Possibilistic MDS).

Conclusions

Perspectives

The same principles could be applied to extend other feature extraction and data visualization methods, such as

- Non linear PCA (e.g. Kernel PCA, principal curves),
- Independent Component Analysis,
- Correspondence analysis,
- etc.

References

- Denœux, T., Masson, M.-H., 2004.
  Principal component analysis of fuzzy data using autoassociative neural networks.
  IEEE Transactions on Fuzzy Systems, 12(3):336–349.
- Hébert, P.-A. , Masson, M.-H., Denœux, T., 2006.
  Fuzzy multidimensional scaling.
  Computational Statistics and Data Analysis, 51(1), 335-359.