IT4Innovations
Centre of Excellence

Dependencies approximated by F–transforms

Iva Tomanová Jiří Kupka

Centre of Excellence IT4Innovations
Division of the University of Ostrava
Institute for Research and Applications of Fuzzy Modeling
Ostrava, Czech Republic

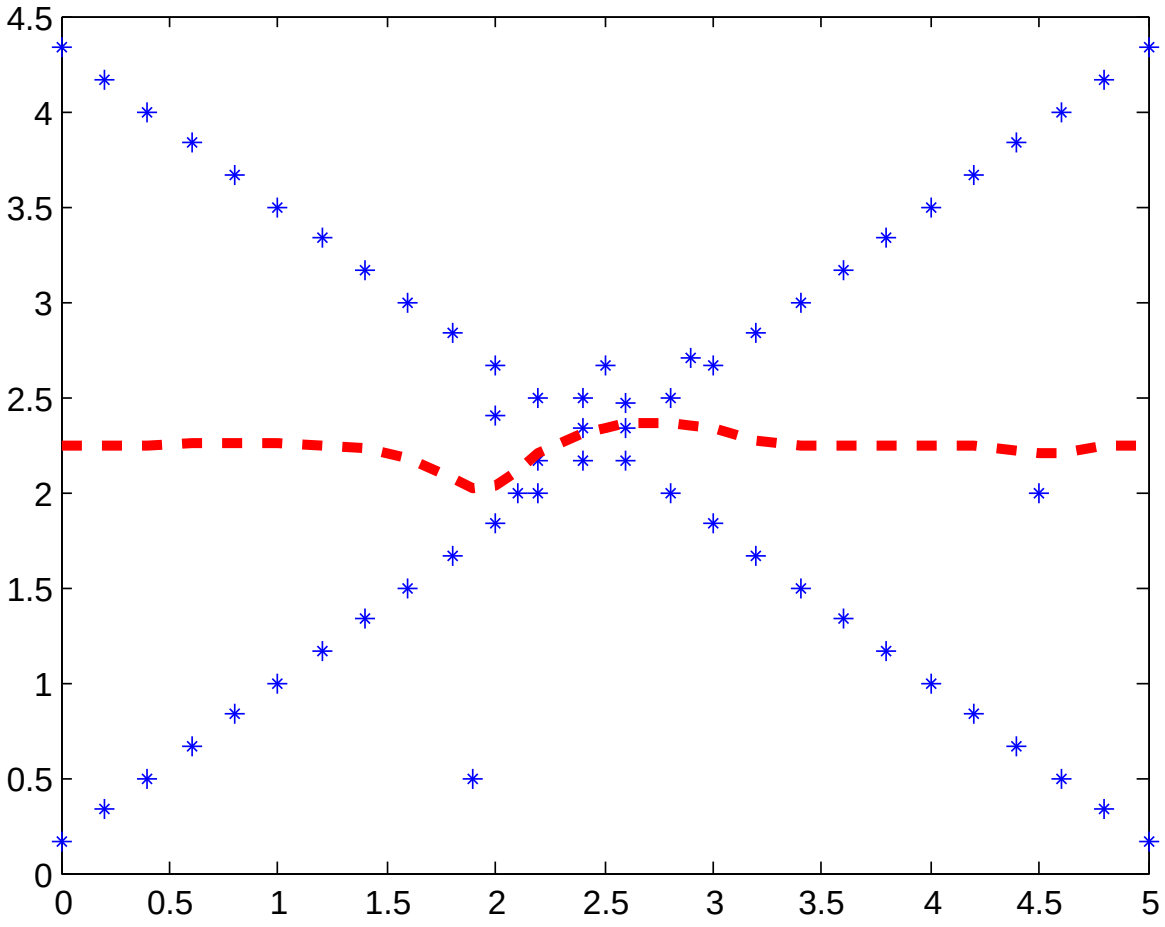
EVROPSKÁ UNIE
EVROPSKÝ FOND PRO REGIONÁLNÍ ROZVOJ
INVESTICE DO VAŠÍ BUDOUCNOSTI

OP Výzkum a vývoj pro inovace

Outline

Outline

Motivation



Outline

Introduction

Data set D

X_1	X_2
e_1	f_1
e_2	f_2
$\vdots$	$\vdots$
e_N	f_N

Fuzzy set

- A on $[a, b] \subseteq \mathbb{R}$,
- is a map $A : [a, b] \to I$.

Outline

DBSCAN algorithm [Ester et al.]

INPUT:

- D . . . a data set,
- $\varepsilon > 0$ . . . a maximal distance between neighboring points in cluster,
- *min_points* . . . a minimal number of points.

OUTPUT:

- The outliers are marked as a noise,
- the number of clusters of arbitrary shape,
- each point either belongs to an appropriate cluster or it is detected as noise.

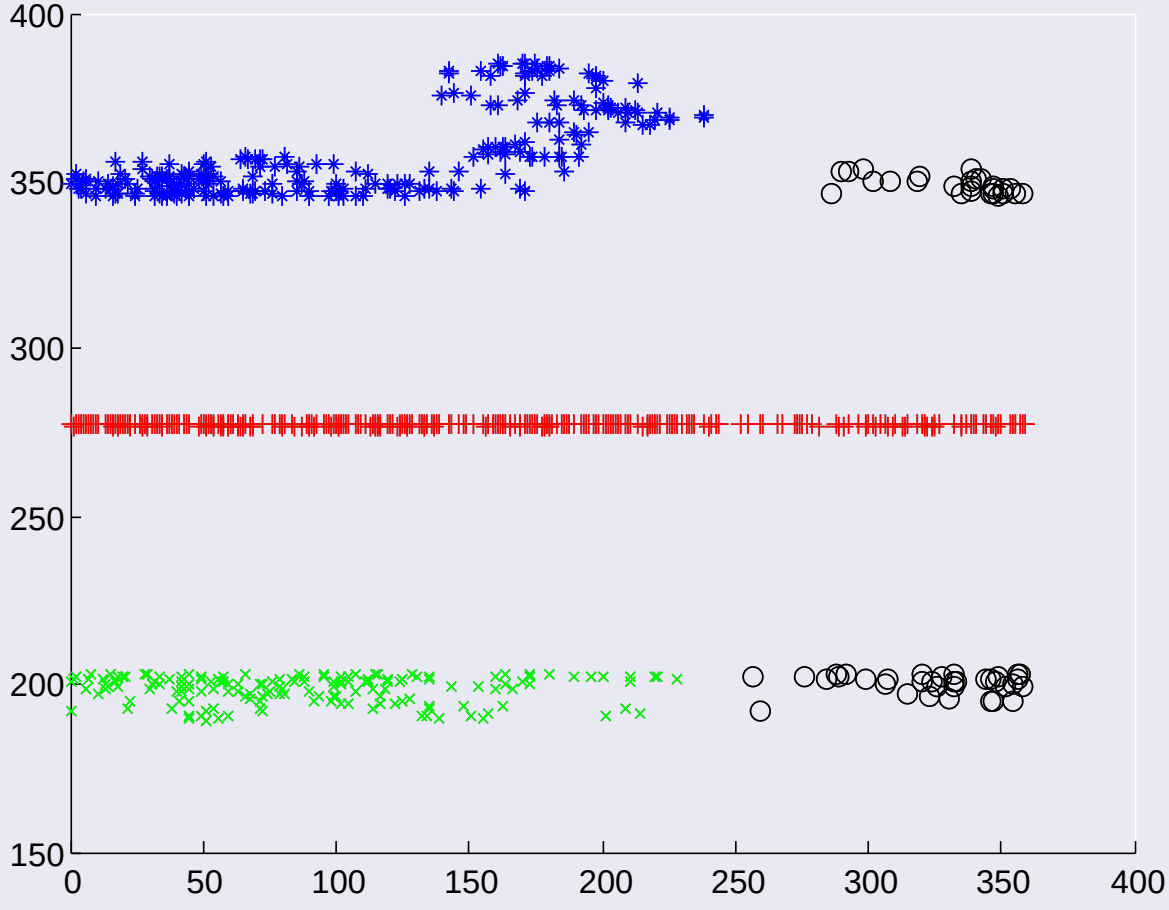
DBSCAN algorithm [Ester et al.]

Main idea:

- To search for objects $z := (e_i, f_i) \in D$ such that the number of its neighbors lying ε-close to z is large enough.
- Such points are then put to the same cluster if they are mutually density-connected, i.e. for any pair $z_1, z_w \in D$ there exists a finite sequence of points $\{z_i\}_{i=1}^{w} \in D$ such that $d_{\mathbb{R}^2}(z_i, z_{i+1}) < \varepsilon$ for any $i = 1, 2, \ldots, w-1$.
- Points of D, which are not ε-close to $z \in C$, are marked as a noise.

DBSCAN

- $\varepsilon = 30$,
- $min\_points = 50$.



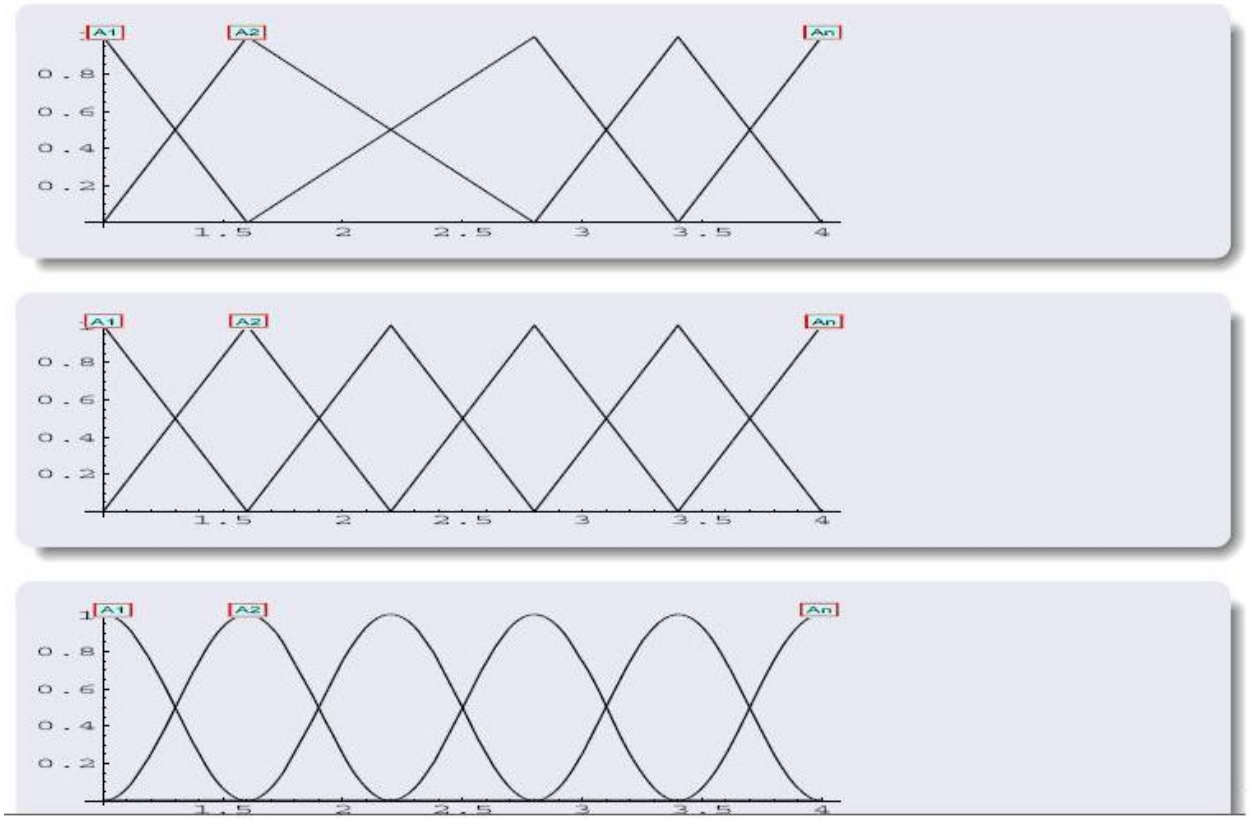
Outline

Discrete F–transform - recall [Perfilieva]

Preliminaries

- Interval $[a, b]$ as a universe,
- $f : [a, b] \to \mathbb{R}$ - given at points $p_0, \ldots, p_N \in [a, b]$,
- $a = x_0 < \cdots < x_n = b, \quad n \geq 3$ - fixed nodes within $[a, b]$,
- $x_k = x_0 + (k-1)h, \quad k = 1, \ldots, n$,
- $h = (b-a)/(n)$ specifies the distance between nodes,
- $A_0, \ldots, A_n : [a, b] \to [0, 1]$ establish an h-uniform fuzzy partition of $[a, b]$.

Fuzzy partitions



Discrete F–transform

Direct discrete F–transform of f

Vector of real numbers $\mathbf{F}_n[f] = [F_0, \ldots, F_n]$ where

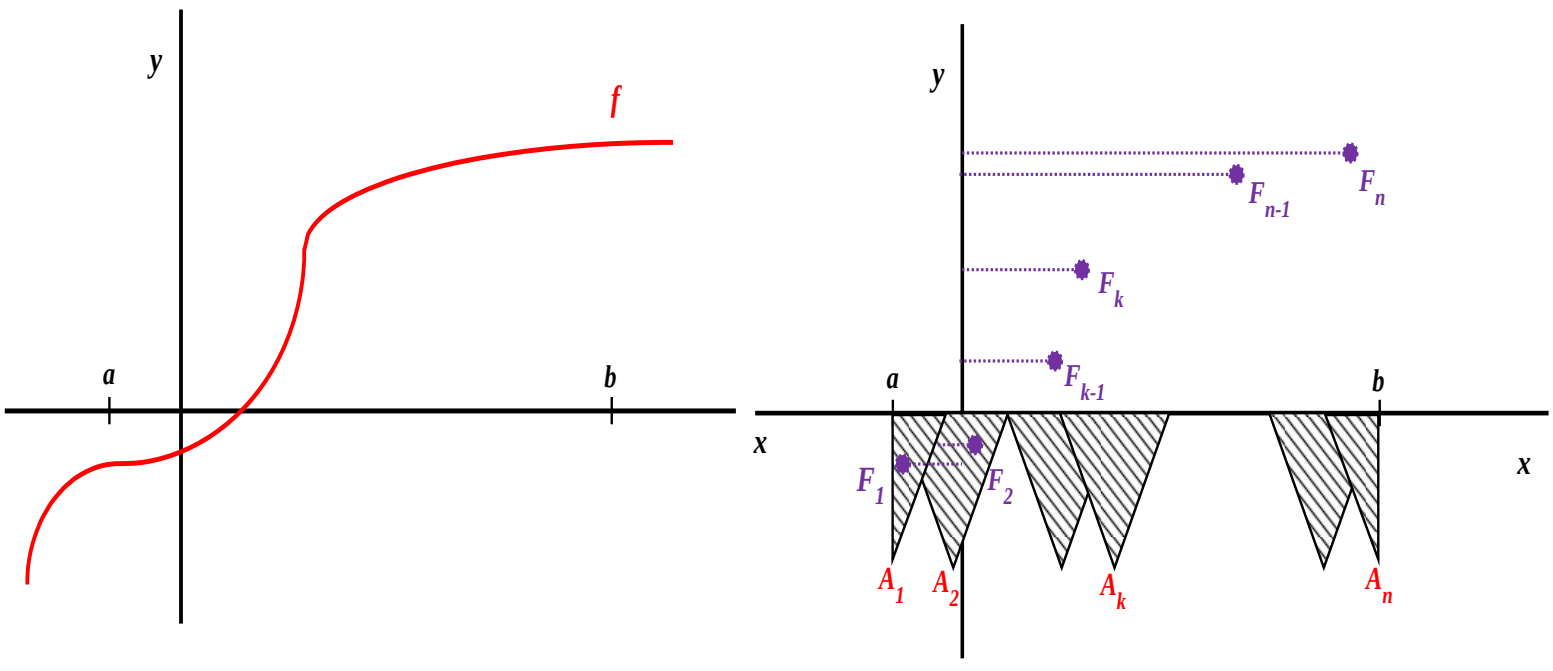
$$F_k = \frac{\sum_{i=1}^{N} A_k(p_i) f(p_i)}{\sum_{i=1}^{N} A_k(p_i)}, \quad k = 0, \ldots, n.$$

Inverse discrete F–transform

$$f_{F,n}(p_i) = \sum_{k=0}^{n} F_k A_k(p_i), \quad i = 1, \ldots, N.$$

F-transform schematically

$$f \longrightarrow [F_0, \ldots, F_n]$$



Theorem 1

Theorem

Let $f(x) : [a, b] \to \mathbb{R}$ be a continuous function. For every $\varepsilon > 0$ there exists an integer $n(\varepsilon)$ and a related fuzzy partition $A_0, A_1, \ldots, A_{n(\varepsilon)}$ of $[a, b]$ such that the inverse F-transform $f_{F,n(\varepsilon)}$ of the function f with respect to the partition $A_0, A_1, \ldots, A_{n(\varepsilon)}$ satisfies

$$|f(x) - f_{F,n(\varepsilon)}(x)| < \varepsilon.$$

Outline

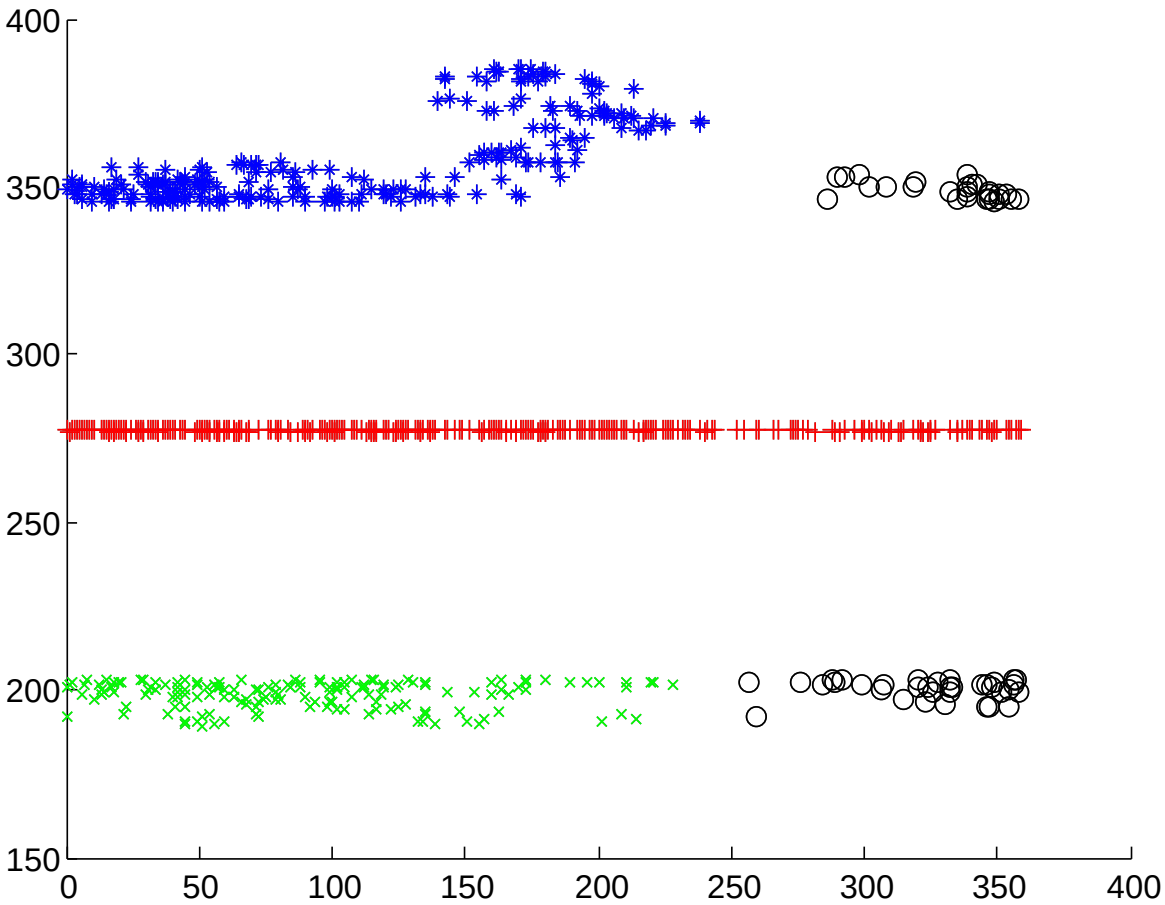
INPUT:

- D . . . a data set,
- *min_points* . . . a minimal number of points in cluster,
- $\varepsilon > 0$ . . . a maximal distance between two neighboring points,
- h . . . a distance between nodes for an h–uniform partition.

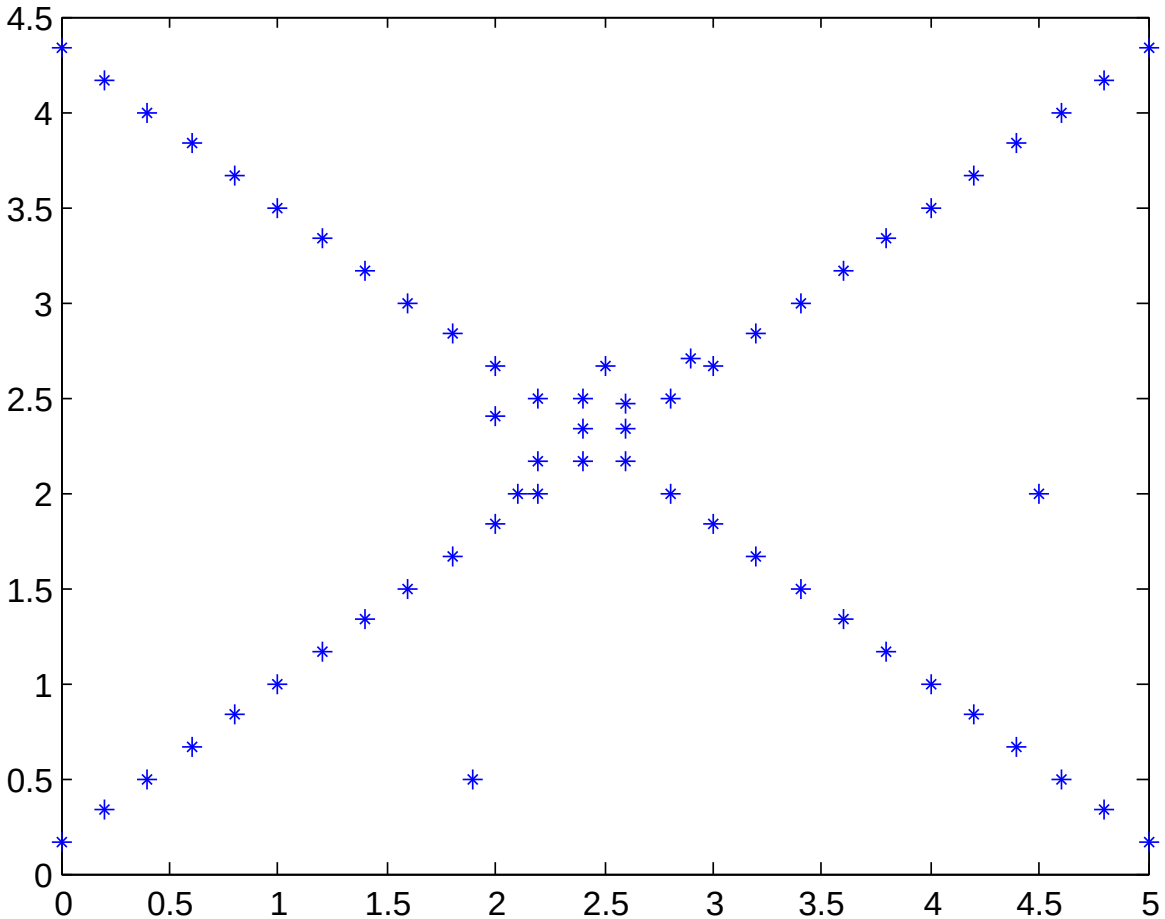
OUTPUT:

- The approximation of any functional dependence with an arbitrary precision,
- the detection of noise (outliers).

1. *DBSCAN* (D, *min_points*, ε) specifies K clusters of D.

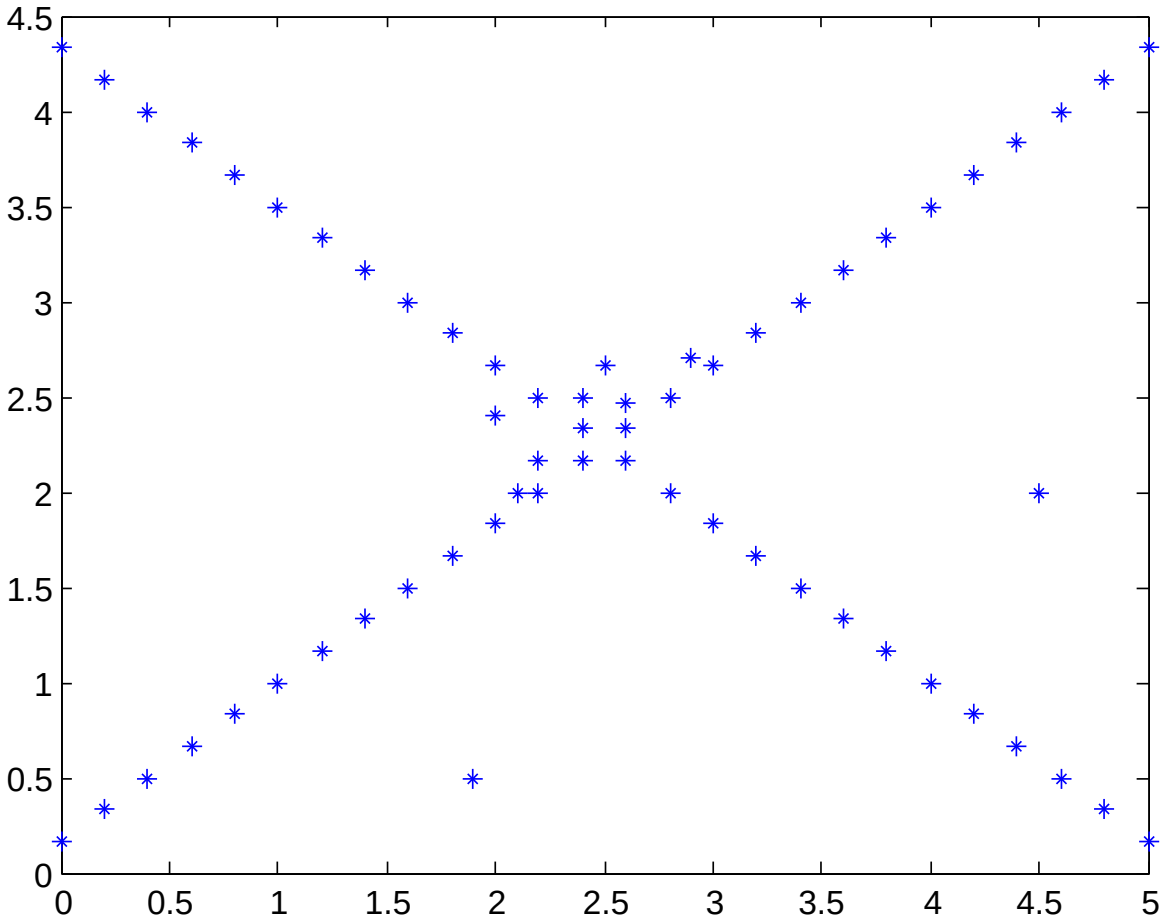

400
350
300
250
200
150
0
50
100
150
200
250
300
350
400

1. *DBSCAN* $(D,\ min\_points,\ \varepsilon)$ specifies K clusters of D.
2. For each cluster C DBSCAN $(C \cap (J_k \times \mathbb{R}), min\_points, \varepsilon)$ checks the appropriateness of C, the sets AC and NC are specified.
   - If $\forall k$ is $K = 1$ then $C \in AC$,
   - otherwise $C \in NC$.

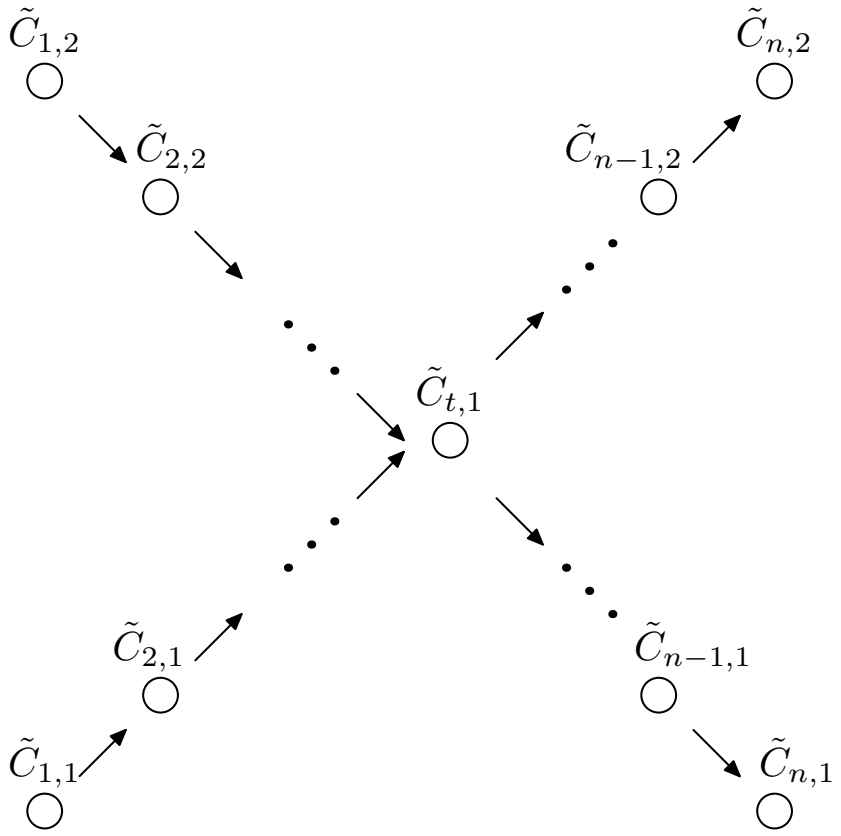

4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

1. *DBSCAN* $(D,\ min\_points,\ \varepsilon)$ specifies K clusters of D.
2. For each cluster C DBSCAN $(C \cap (J_k \times \mathbb{R}), min\_points, \varepsilon)$ checks the appropriateness of C, the sets AC and NC are specified.
3. For every cluster $C \in NC$ maximal oriented paths are constructed.

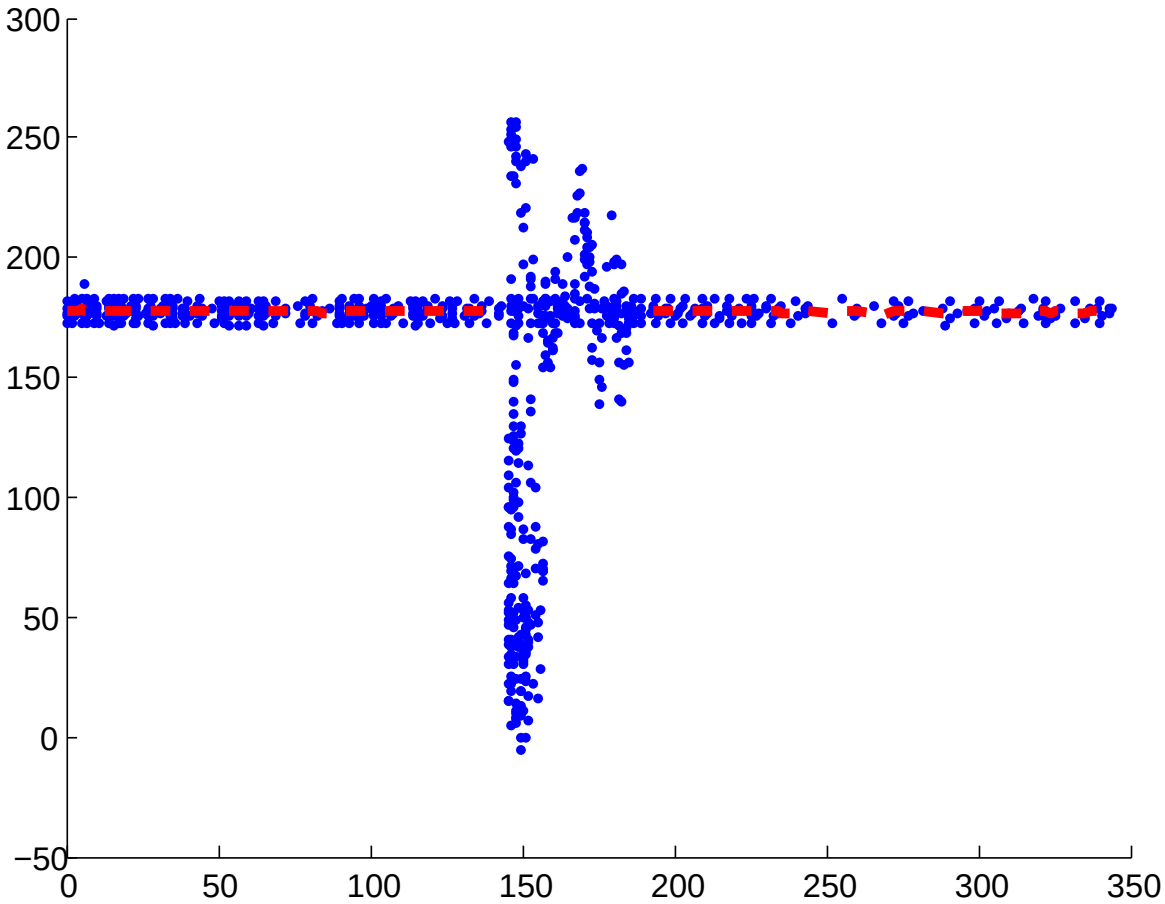
Illustrative cluster $C \in NC$



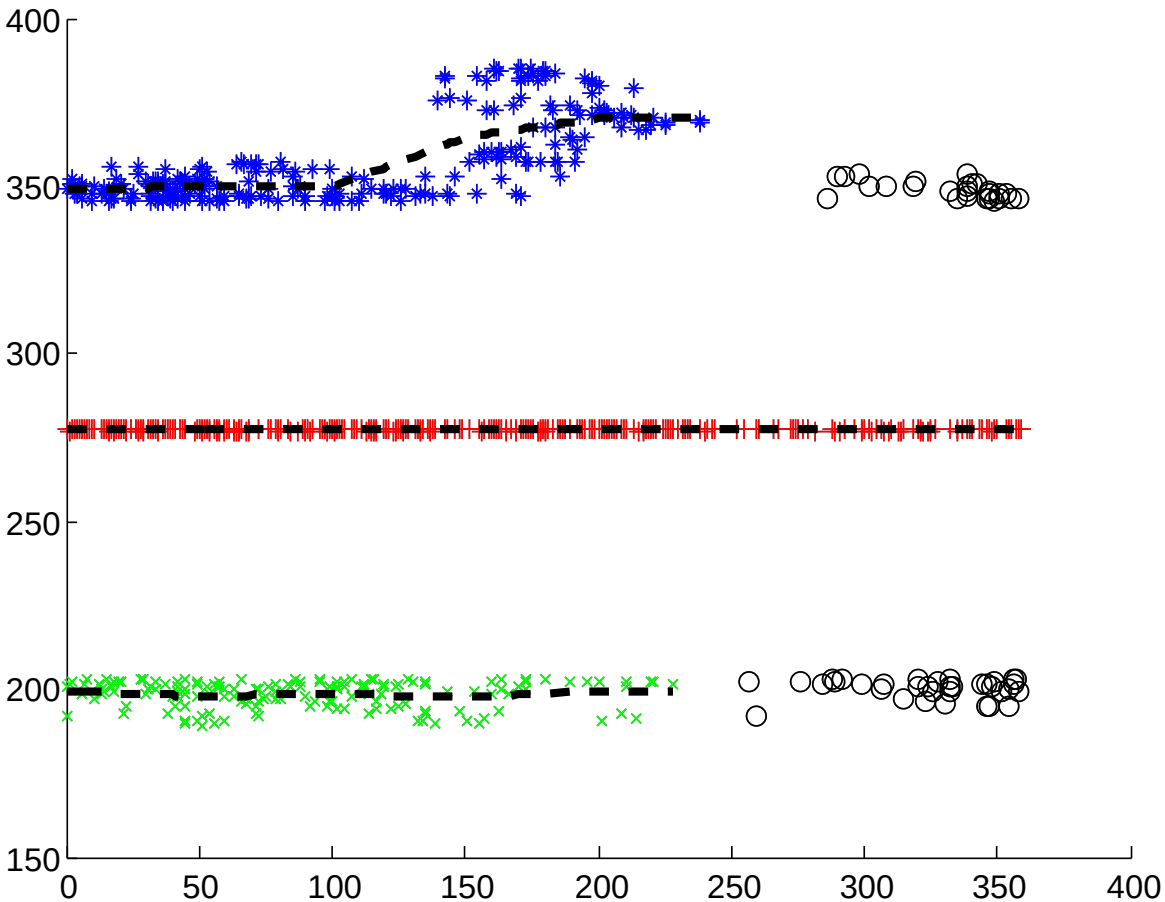
Oriented graph representation of the cluster C



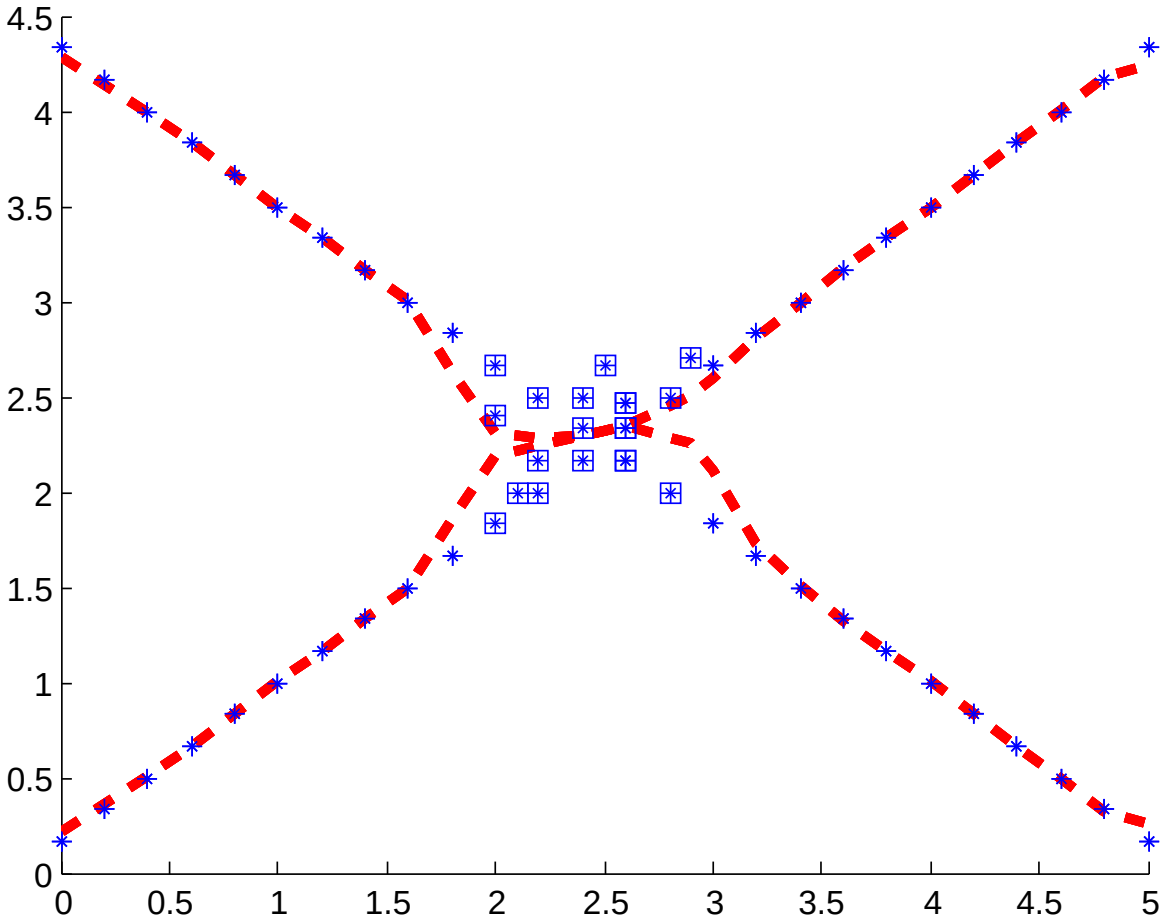
1. *DBSCAN* (D, *min_points*, ε) specifies K clusters of D.
2. For each cluster C DBSCAN ($C \cap (J_k \times \mathbb{R})$, *min_points*, ε) checks the appropriateness of C, the sets AC and NC are specified.
3. For every cluster $C \in NC$ maximal oriented paths are constructed.
4. For any $C \in AC$ and for any $C \in ASC_i$ of $C \in NC$,
   - check the accuracy of C, (if necessary, specify a part of D which cannot be approximated by a continuous function) modify C if necessary,
   - approximate C by the inverse F-transform.


300
250
200
150
100
50
0
−50
0
50
100
150
200
250
300
350

Example 1



Example 2



Theorem 2

Theorem

Let D be a subset of $\mathbb{R}^2$ which can be expressed as the union of finitely many continuous functions f_j on $[a, b]$. Then for every $\varepsilon > 0$ there exists an integer $n(\varepsilon)$, a related fuzzy partition $A_0, A_1, \ldots, A_{n(\varepsilon)}$ of $[a, b]$ and such that the inverse F–transforms of F–transforms given by

$$\tilde{\mathbf{F}}[f] = [\tilde{F}_0, \ldots, \tilde{F}_n]$$

are ε-close to the graphs of f_j's, where $\tilde{F}_i$ is a finite number of i-th components F_i of F–transforms calculated for some appropriate clusters or subclusters of D.

Outline

Conclusions

Our algorithm

- is a simple procedure expanding applicability of the technique of F–transform in data analysis,
- is a preprocessing step as well as automatically provides several advanced features,
- provides a fast outliers detection,
- detects all functional dependencies which can be represented by a finitely many continuous functions,
- detects parts of data set without functional dependencies.

Thanksgiving

Thank You for Your Attention