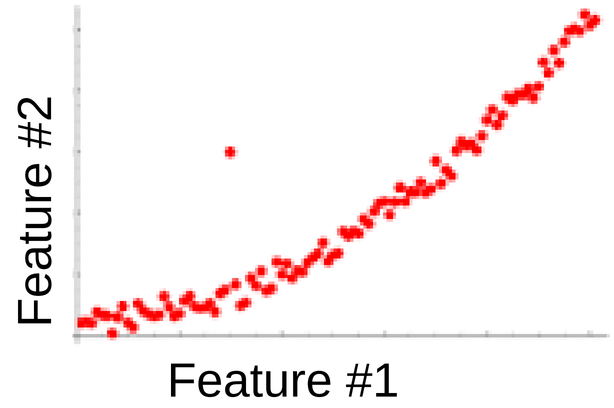


Network tools for outlier detection

Cristina Masoller

Departamento de Fisica

Universitat Politècnica de Catalunya

Complexity in Economics & Finance
StatPhys29 Satellite, Ministry of Internal Affairs
Rome, Italy, July 8, 2025

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Campus d'Excel·lència Internacional

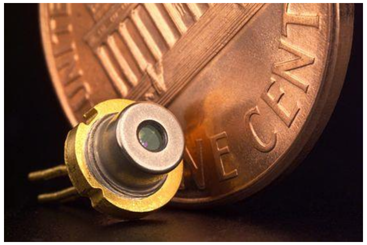
ICREA

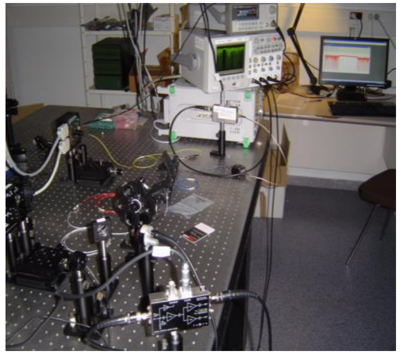
AGAUR

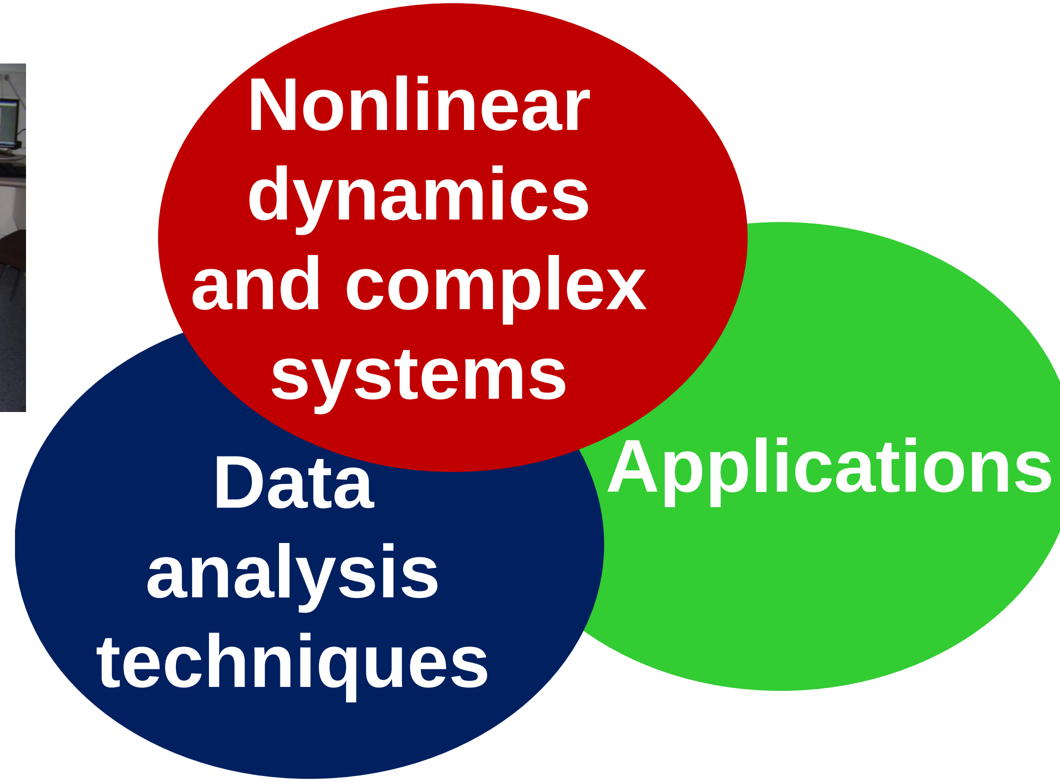
GOBIERNO DE ESPAÑA

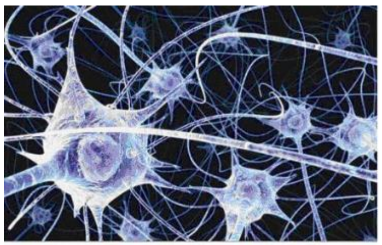
MINISTERIO DE CIENCIA E INNOVACIÓN

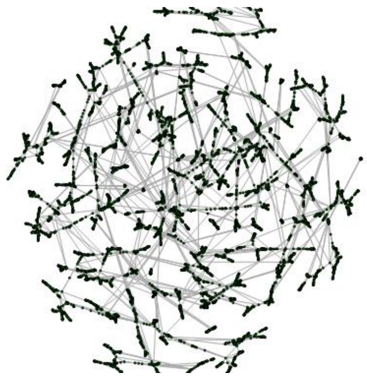
Research lines

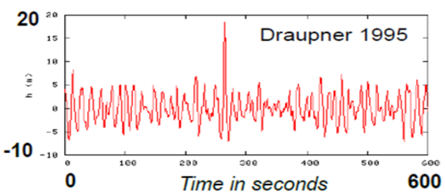


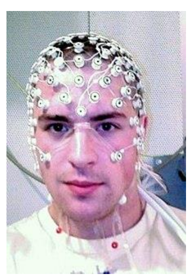


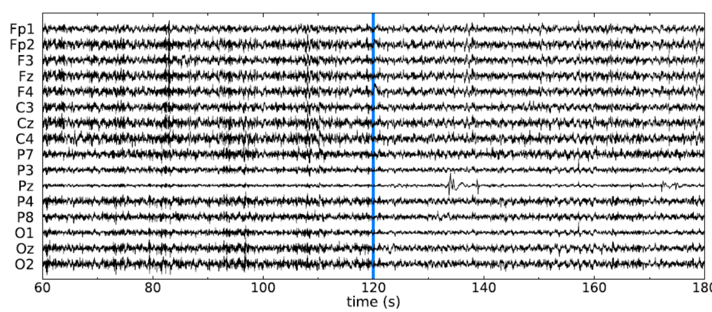


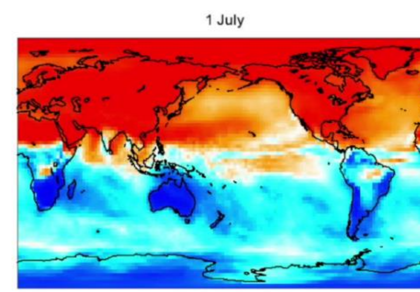












cristina.masoller@upc.edu @cristinamasoll1

Strange observation/data: an outlier or an anomaly/artifact?

- *Anomaly*: a data point that cannot be explained given current knowledge of the process that generates the data.

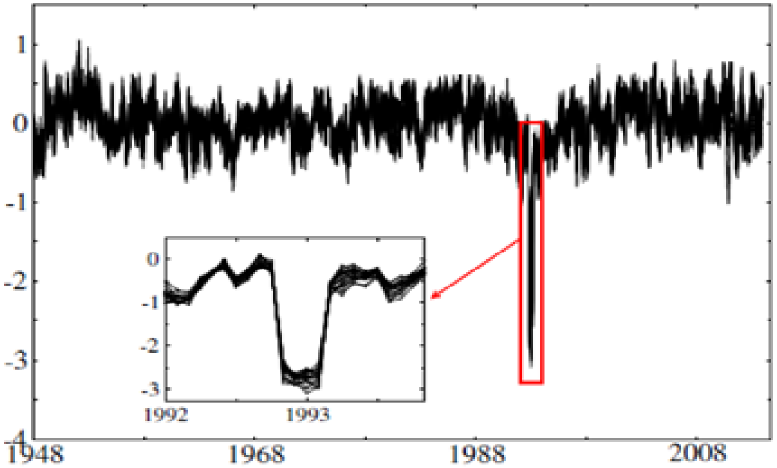
- *Outlier*: a “legitimate” data point that is far from the center of the distribution that characterizes the process.

- *Novelty* detection: “new” event/data not seen before.

cristina.masoller@upc.edu @cristinamasoll1

Types

- *Point anomalies*: a data point that is anomalous with respect to the rest of the data.
- *Contextual anomalies*: a data point that is anomalous in a specific context.
- *Collective anomalies*: a set of data points that are not anomalies by themselves, but their collective occurrence is anomalous.

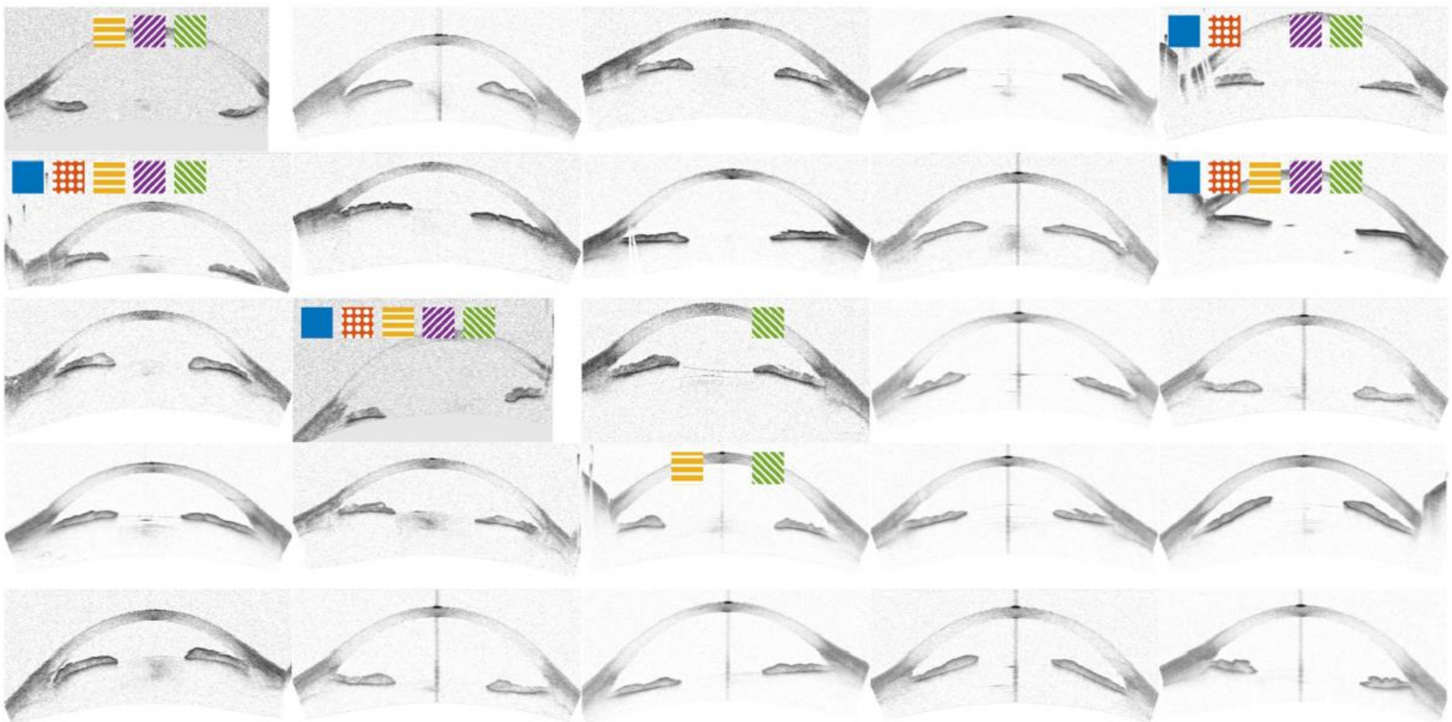


V. Chandola et al., ACM Comput. Surveys 41, 15 (2009)

cristina.masoller@upc.edu @cristinamasoll1

“Practical” definition:

Removing outliers / anomalies from the training data improves the performance of a machine learning algorithm.

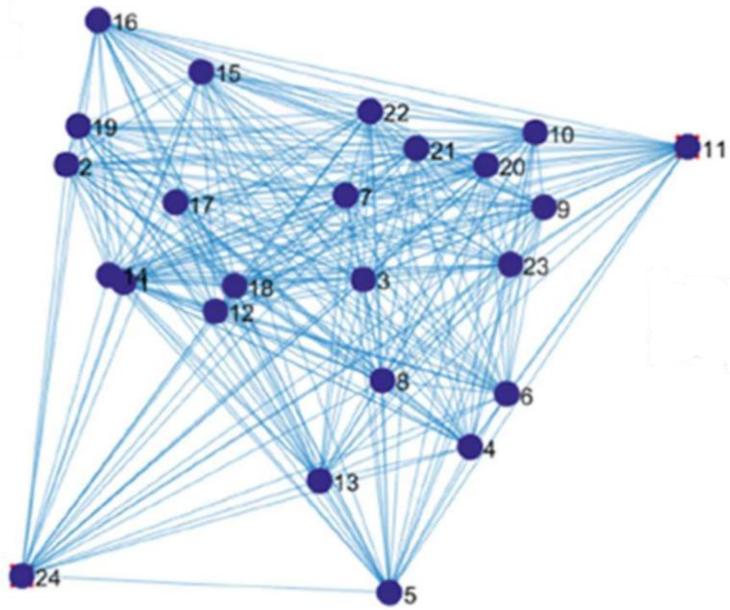


cristina.masoller@upc.edu @cristinamasoll1

We consider a dataset of high-dimensional items where a distance can be defined between pairs of items.

Feature vectors of items *i* and *j*: $\{f_{i1} \ldots f_{iM}\}$ $\{f_{j1} \ldots f_{jM}\}$

Euclidian distance: $$d_{ij} = \sqrt{\sum_{k=1}^{M}(f_{ik} - f_{jk})^2}$$



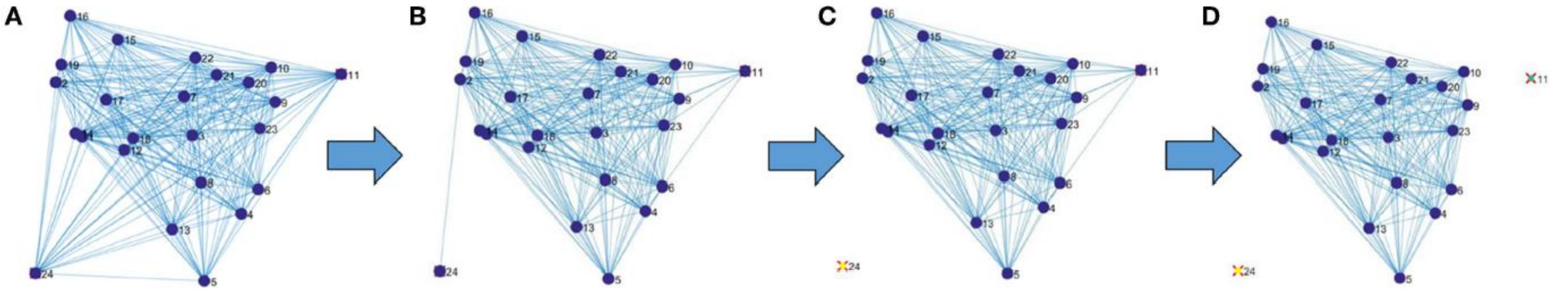
Fully connected network

The weights of the links are the distances

cristina.masoller@upc.edu @cristinamasoll1

First method: Outlier detection using **percolation**

Sequentially remove links with longest distances

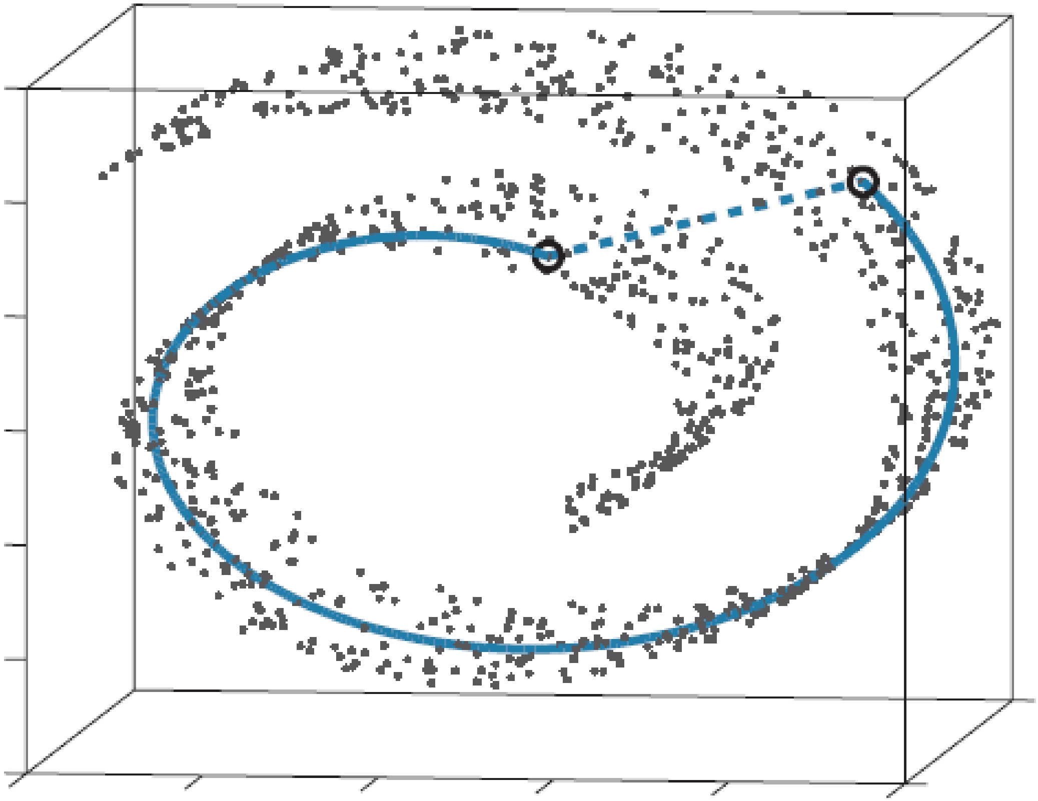


Outlier score of an item = order in which the item disconnects from the giant component. **Parameter free.**

cristina.masoller@upc.edu @cristinamasoll1

Second method: nonlinear dimensionality reduction

- *Main idea: how well or how poorly an element fits in the learned manifold*.



Distance in the high dimensional space (dash) and distance in the learned (lower dimensional) manifold (solid).

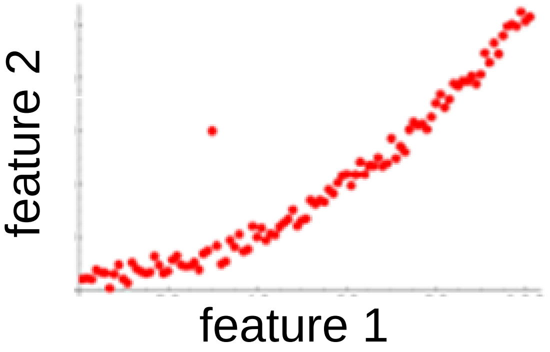
IsoMap, Tenenbaum et al., Science 290, 2319 (2000).
P. Amil, N. Almeida and C. Masoller, Front. Phys. 7, 194 (2019).

cristina.masoller@upc.edu @cristinamasoll1

Steps

- Apply *IsoMap* to the distance matrix $\mathbf{D_{ij}}$ to obtain
  - a new set of features
  - a new distance matrix in the geodesic space, $\mathbf{D^G_{ij}}$
- With the new features, recalculate the distance matrix $\mathbf{D'_{ij}}$
- For each element, calculate correlation, between $\mathbf{D^G_{ij}}$ and $\mathbf{D'_{ij}}$
- <u>Outlier score</u>: $OS_i = 1-\rho_i^2$
- Two parameters (integers):
  - Dimension of reduced space
  - # of geodesic neighbors

Note: we don't use the features returned by IsoMap to assign outlier scores



cristina.masoller@upc.edu @cristinamasoll1

Comparison with other distance-based outlier mining methods

- Distance to center of mass (d2CM): an outlier score is assigned according to the distance of an element to the center of mass.
- Ramaswamy: an outlier score is assigned according to the distance of an element to its kth nearest neighbor.
- One Class Support Vector Machine (OCSVM): uses the scalar product to define a function that returns +1 in the region where normal elements are located and −1 elsewhere.

Ramaswamy et al., Efficient algorithms for mining outliers from large data sets. ACM Sigmod Record (Vol. 29, No. 2, pp. 427-438, 2000).
Schölkopf et al., Estimating the support of a high-dimensional distribution. Neural computation13, 1443-1471, 2001.

cristina.masoller@upc.edu @cristinamasoll1

Face database

We added to some random images a square with gray-scale pixels whose color distribution is the same as that of the image.

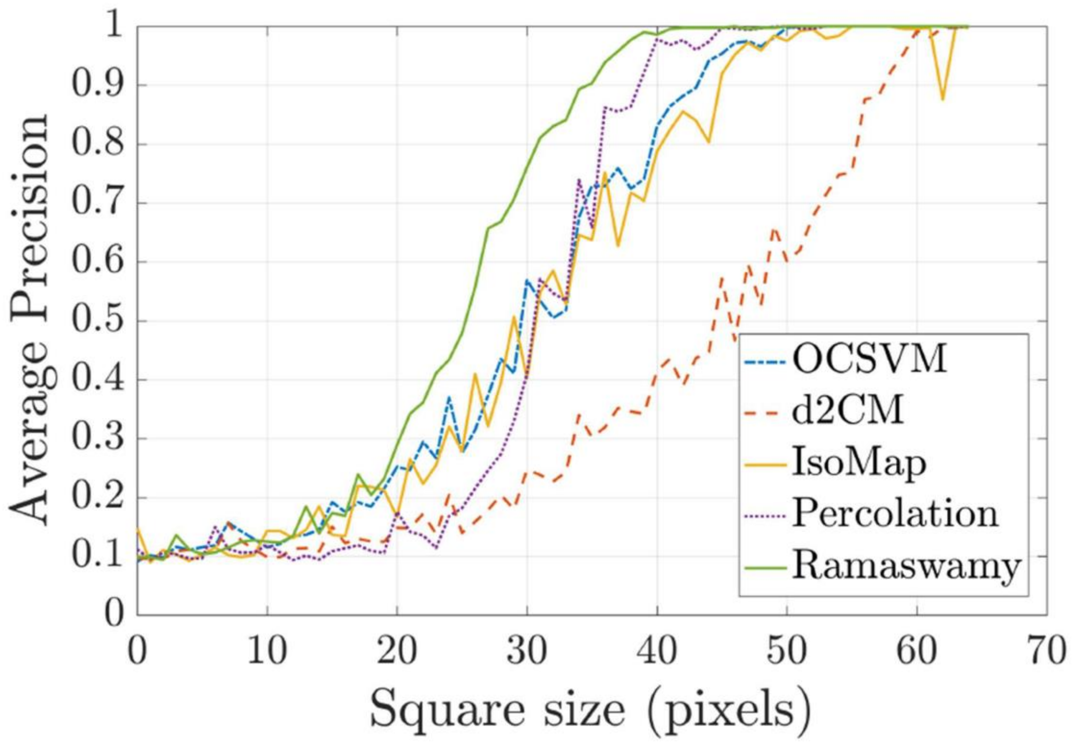
http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

cristina.masoller@upc.edu @cristinamasoll1

Performance quantification:

Average precision: area under the Precision-Recall curve, TP/(TP+FP) vs TP

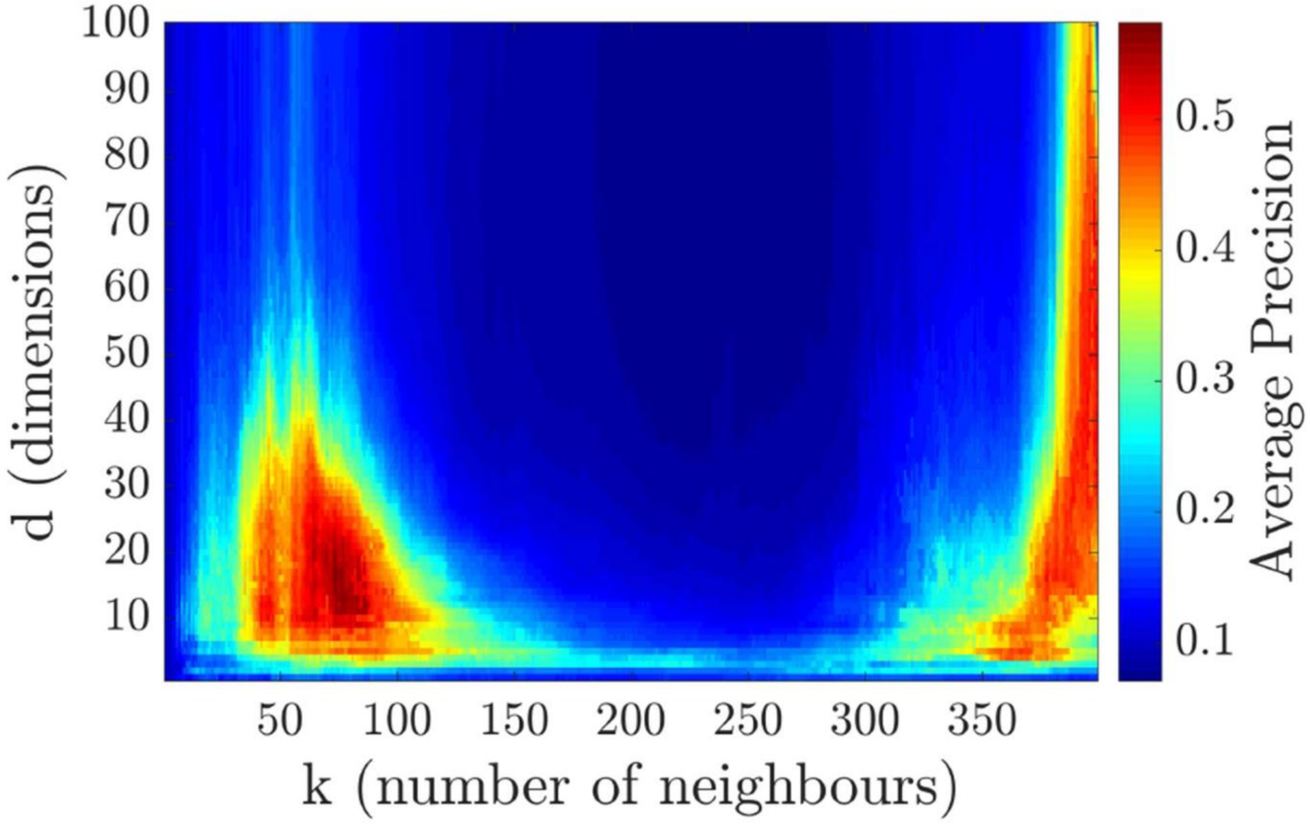
It does not depend on the number of true negatives.



P. Amil, N. Almeida and C. Masoller, Front. Phys. 7, 194 (2019).

cristina.masoller@upc.edu @cristinamasoll1

ISOMAP precision can be improved by selecting the parameters

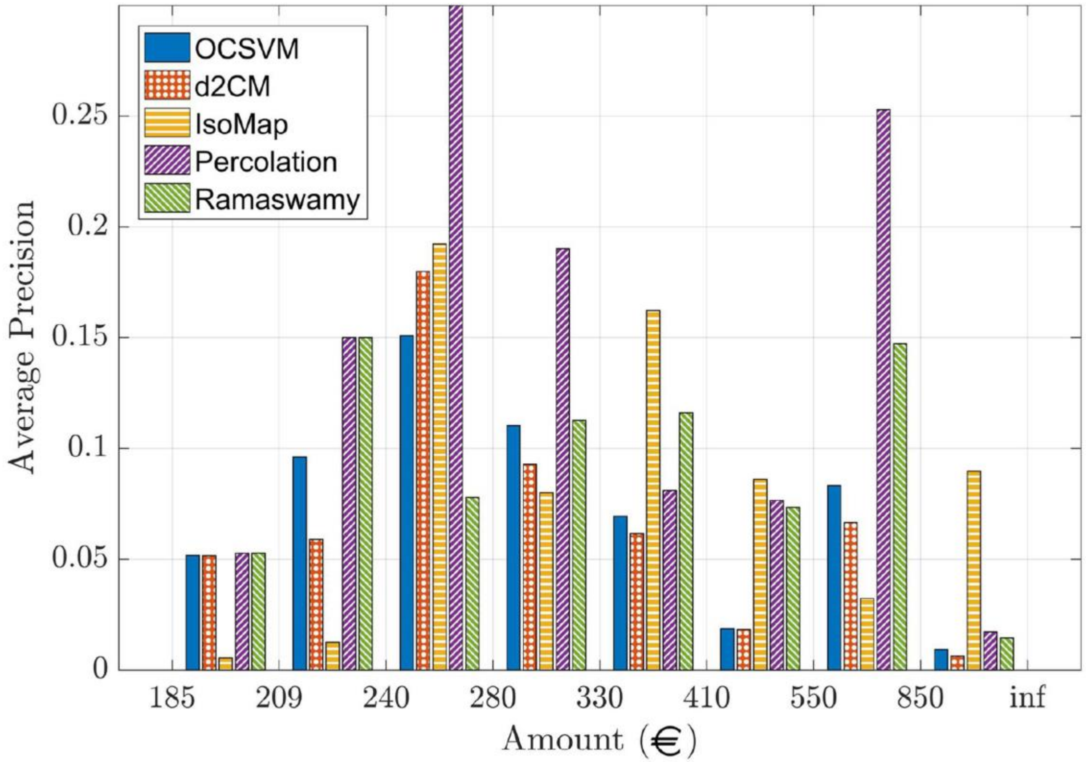


square size: 30 pixels

cristina.masoller@upc.edu @cristinamasoll1

How about other types high-dimensional of items?

We analyzed a database of Credit Card Transactions (some labeled as frauds); each transaction has 28 features from PCA.



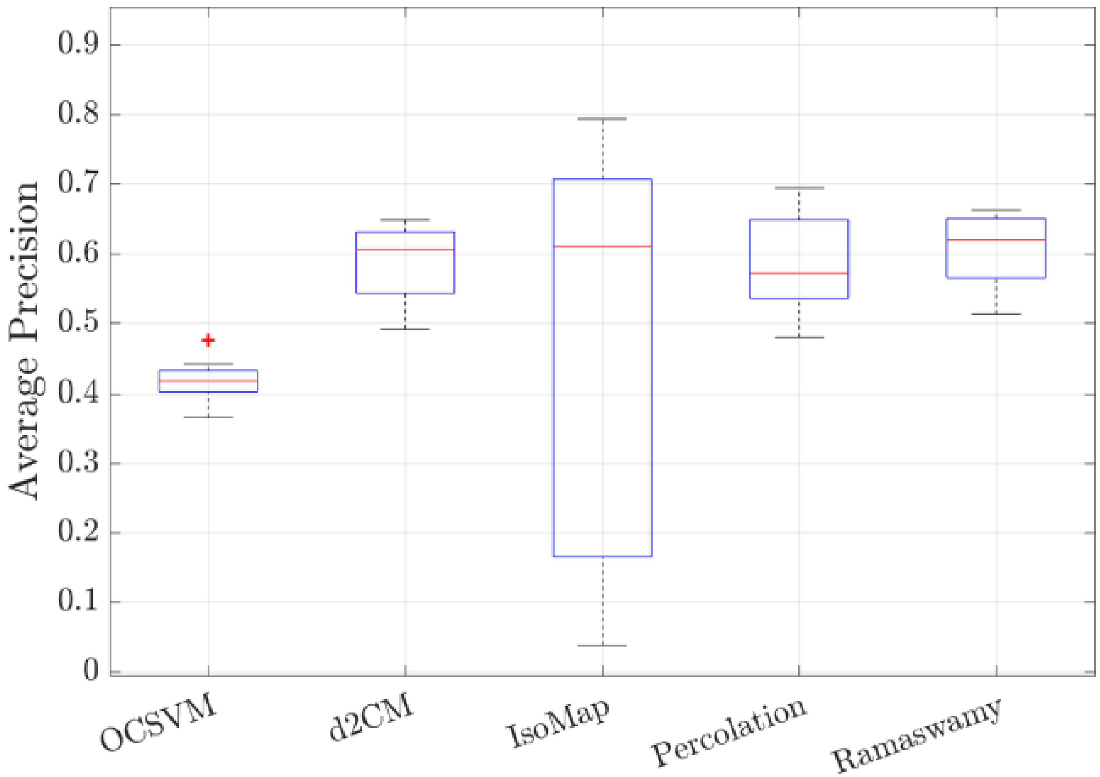
4,000 transactions
100 labeled as frauds.
2000 for training and 2000 for testing.

https://www.kaggle.com/mlg-ulb/creditcardfraud

P. Amil, N. Almeida and C. Masoller, Front. Phys. 7, 194 (2019).

cristina.masoller@upc.edu **@cristinamasoll1**

Analysis of 7 sets of 4000 credit transactions, chosen without considering the amount of the transaction.
In each set: 3900 regular and 100 frauds



(reminder) Average precision: area under the Precision-Recall curve, TP/(TP+FP) vs TP

https://www.kaggle.com/mlg-ulb/creditcardfraud

P. Amil, N. Almeida and C. Masoller, Front. Phys. 7, 194 (2019).

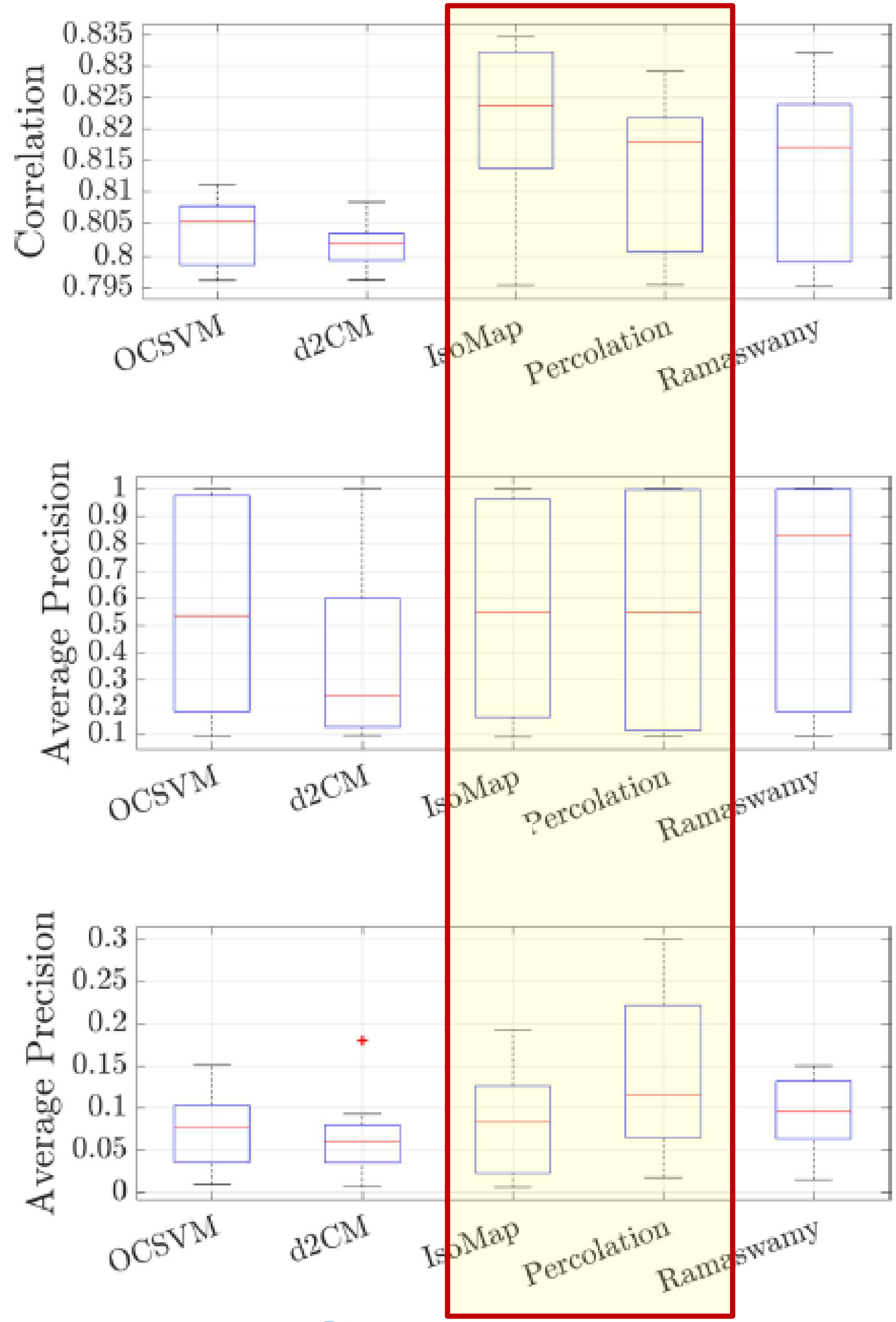
cristina.masoller@upc.edu @cristinamasoll1

Summary of results

OCT images

Face images

Credit card transactions



cristina.masoller@upc.edu @cristinamasoll1

The methods' performance depends on the data.
How do they compare in terms of execution time?

For a database of 1,000 elements with 30 dimensions, run on *Matlab* on an Intel i7-7700HQ laptop:

Distance to center of mass	0.01 s
Ramaswamy	0.04 s
One Class Support Vector Machine	0.2 s
Percolation	6 s
IsoMap	18 s

Can we do better?

cristina.masoller@upc.edu @cristinamasoll1

Another distance-based way to mine anomalies / outliers

Feature vectors of items *i* and *j*

$$\{f_{i1} \dots f_{iM}\} \qquad \{f_{j1} \dots f_{jM}\}$$

1. Euclidean distance: $d_{ij} = \sqrt{\sum_{k=1}^{M} (f_{ik} - f_{jk})^2}$ $\boxed{OS1_i = \frac{1}{N}\sum_{l=1}^{N} d_{il}}$

2. Jensen-Shannon divergence: distance between the distributions of distances of items "j" and "k": $P_j=\{d_{jl}\}$ & $Q_k=\{d_{kl}\}$

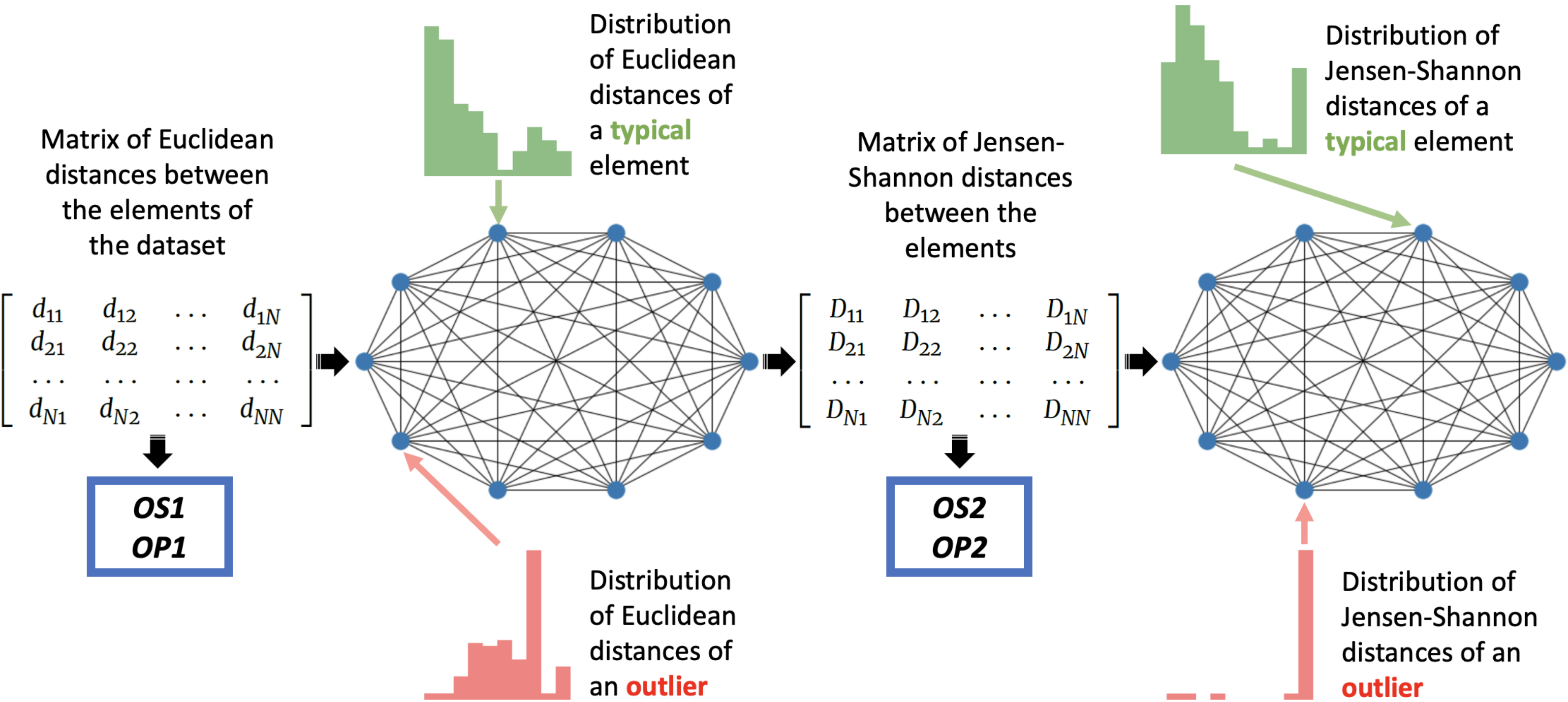
$$D_{jk} = \frac{1}{2}\left[\sum_{i=1}^{d} P_i \ln\left(\frac{2P_i}{P_i+Q_i}\right) + \sum_{i=1}^{d} Q_i \ln\left(\frac{2Q_i}{P_i+Q_i}\right)\right]$$

$$\boxed{OS2_i = \frac{1}{N}\sum_{l=1}^{N} D_{il}}$$

First case: the weights are the distances between feature vectors. Second case: weights are distances between probability distributions.

In both cases: outlier score is the sum of the weights of the links of the node.

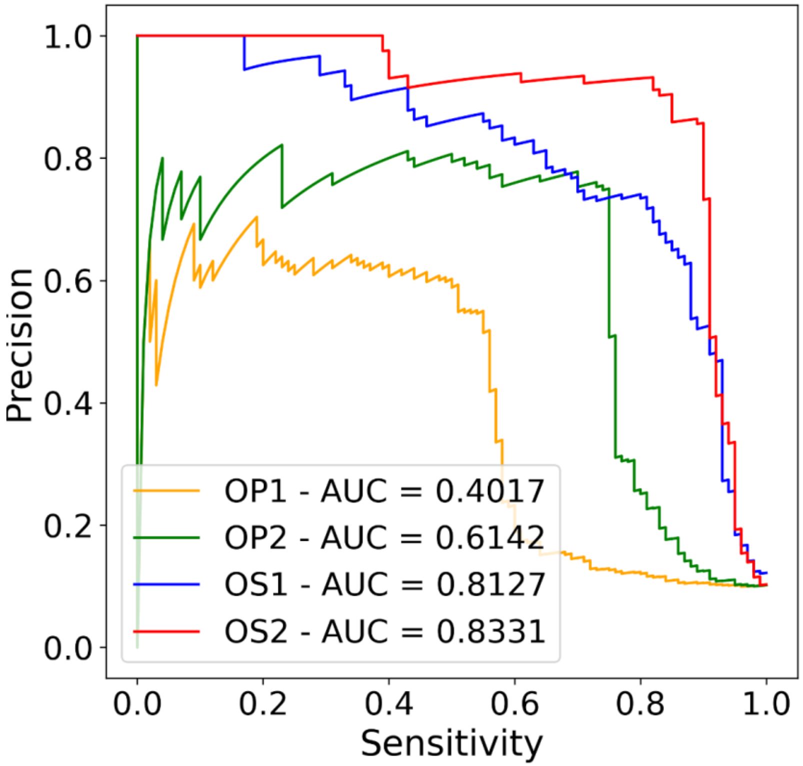
cristina.masoller@upc.edu @cristinamasoll1

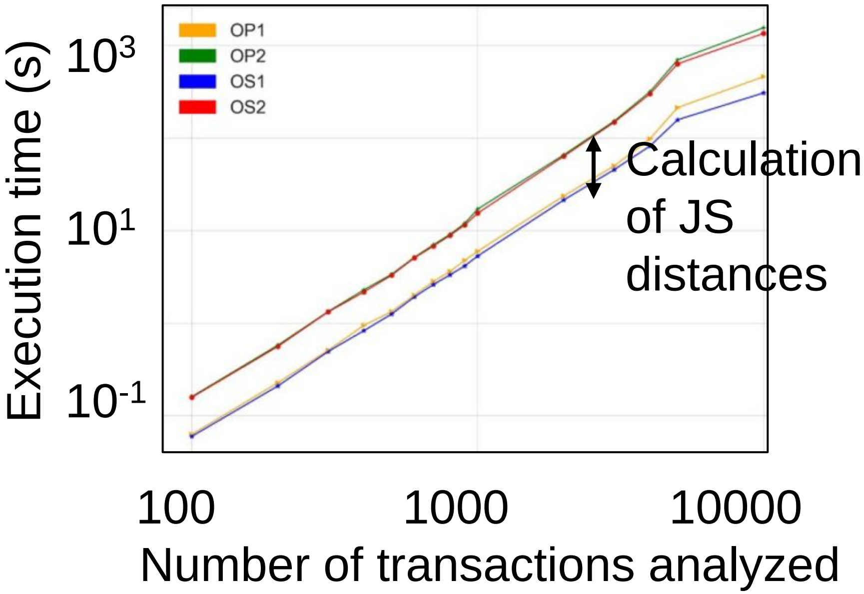


OS1, OS2: Outlier score defined by the sum of distances
OP1, OP2: Outlier score defined by the percolation order

cristina.masoller@upc.edu @cristinamasoll1

Results 1000 credit card transactions: 10% fraud (similar with 5% fraud)





OP1: Percolation, Euclidian distances
OS1: Sum of Euclidian distances

OP2: Percolation, JS distances
OS2: Sum of JS distances

A. S. O. Toledo et. al, *Outlier mining in high-dimensional data using the Jensen-Shannon divergence and graph structure analysis*, J. of Phys: Complexity 3, 045011 (2022).

cristina.masoller@upc.edu @cristinamasoll1

Average precision for different sizes (90% normal transactions, 10% frauds)

Table 1. Performance obtained for datasets of different sizes, N. For each N the mean and the standard deviation of the AUC-PR were calculated from 200 datasets composed by different elements, such that 90% are normal transactions and 10% are frauds.

	OP1		OP2		OS1		OS2	
N	mean	std	mean	std	mean	std	mean	std
100	0.53	0.17	0.48	0.11	0.82	0.11	0.85	0.10
200	0.48	0.12	0.55	0.12	0.82	0.08	0.85	0.07
500	0.41	0.08	0.61	0.08	0.82	0.05	0.84	0.05
1000	0.40	0.07	0.61	0.07	0.81	0.04	0.83	0.04
5000	0.38	0.04	0.62	0.04	0.81	0.02	0.82	0.02
10 000	0.37	0.03	0.62	0.03	0.82	0.02	0.83	0.02

(Reminder)

Average precision: area under the Precision-Recall curve, TP/(TP+FP) vs TP

cristina.masoller@upc.edu @cristinamasoll1

Take home messages

The outlier mining methods proposed require to define a meaningful distance between the elements of a database that have associated high-dimensional "feature" vectors.

Parameter free.

The database can not be too large because the execution time grows at least as NxN with the size N of the database (but linearly with the number of features).

Can be used to mine outliers in time-series data, two-dimensional data (images), unstructured data, etc.

cristina.masoller@upc.edu @cristinamasoll1

The doers: Pablo Amil & Alex Toledo

References:

- P. Amil et al., "*Outlier mining methods based on network structure analysis*", Front. Phys. 7, 194 (2019).
- A. S. O. Toledo et. al, "*Outlier mining in high-dimensional data using the Jensen-Shannon divergence and graph structure analysis"*, J. of Phys: Complexity 3, 045011 (2022).

A. S. O. Toledo et. al, "*Outlier mining in criminal networks: The role of machine learning and outlier detection models*"
J. Comput. Soc. Sci. 8, 36 (2025).

Thank you for your attention !

cristina.masoller@upc.edu @cristinamasoll1