Anomaly detection in hidden subspaces

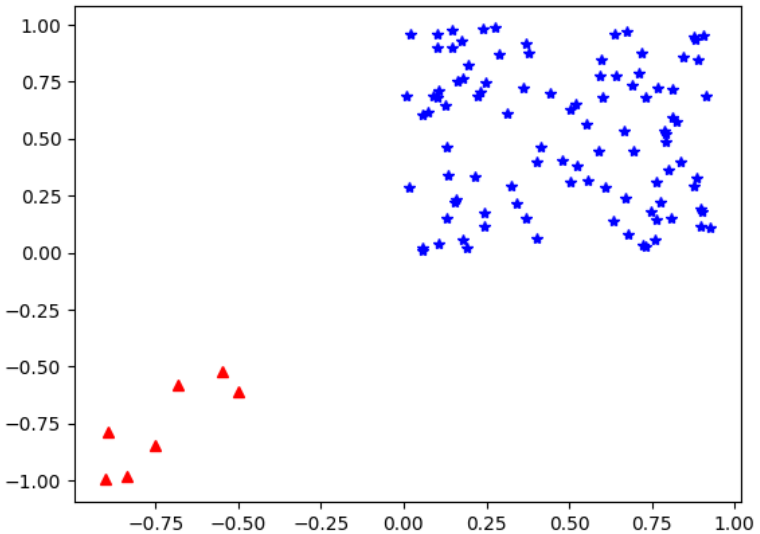
Luis A. Souto Arias

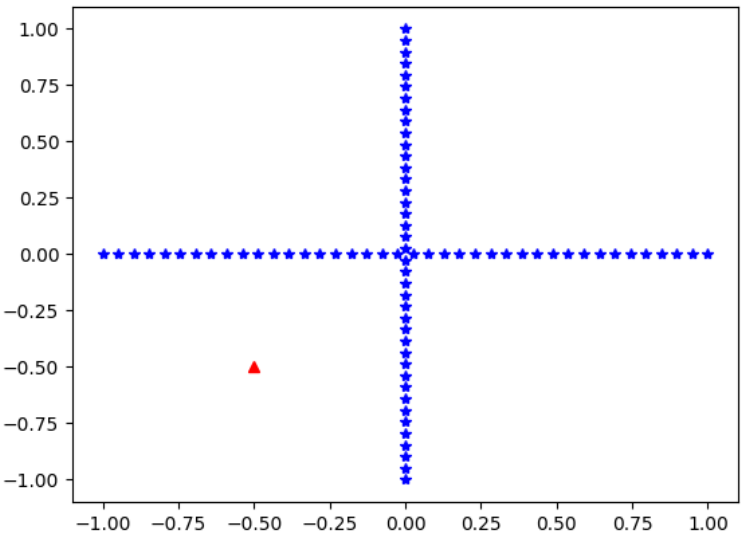
What is an anomaly/outlier?

Definition (Hawkins 1980):

"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism."

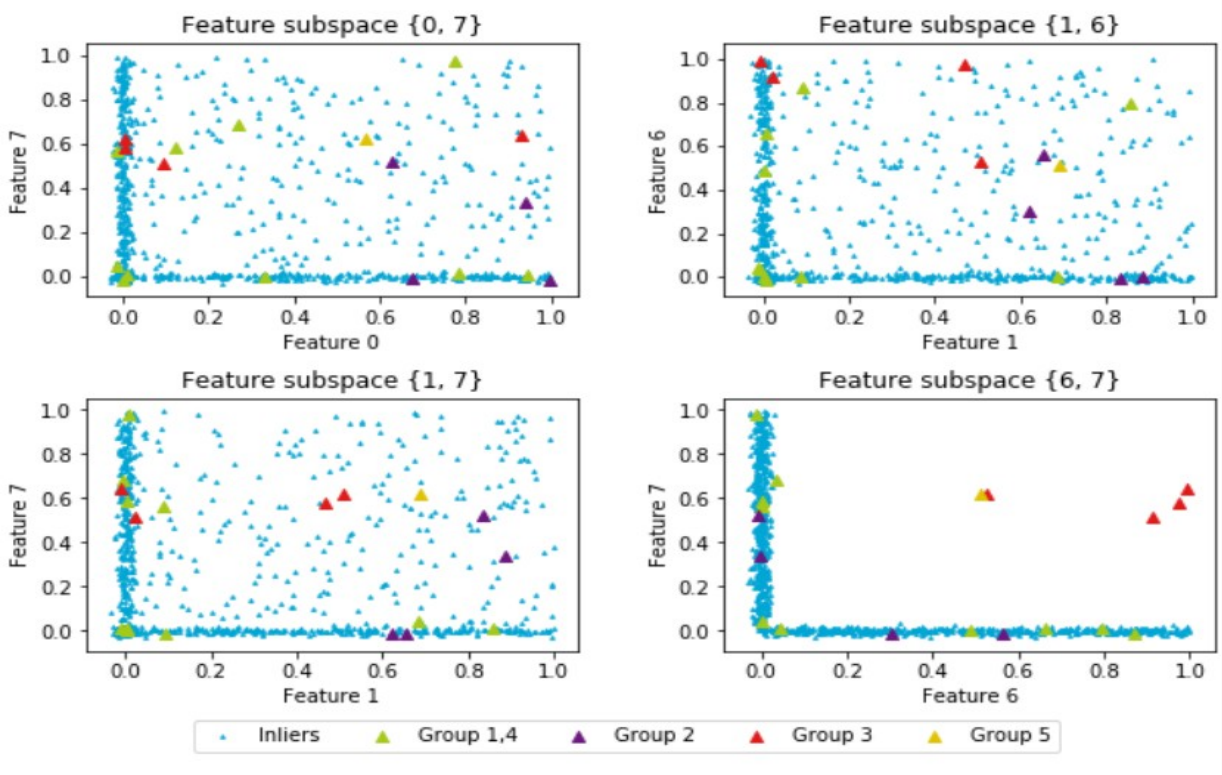
Examples of outliers





Hidden subspace

HiCS example: Data set of 1000 observations and 10 features, with 19 outliers hidden in two-dimensional subspaces.



Standard approaches

Two of the most popular methods for anomaly detection are k-Nearest Neighbours (kNN) and Isolation Forest (IF).

The former is based on the concept of **distance** while the latter is based on the concept of **isolation**.

k-Nearest Neighbours

As the name implies, kNN computes an anomaly score that is based on the distance of each point to it kNNs.

Depending on the value of k, it may not detect clusters of outliers.

Computation of pairwise distances implies a computational complexity of O(n log(n)).

Isolation Forest

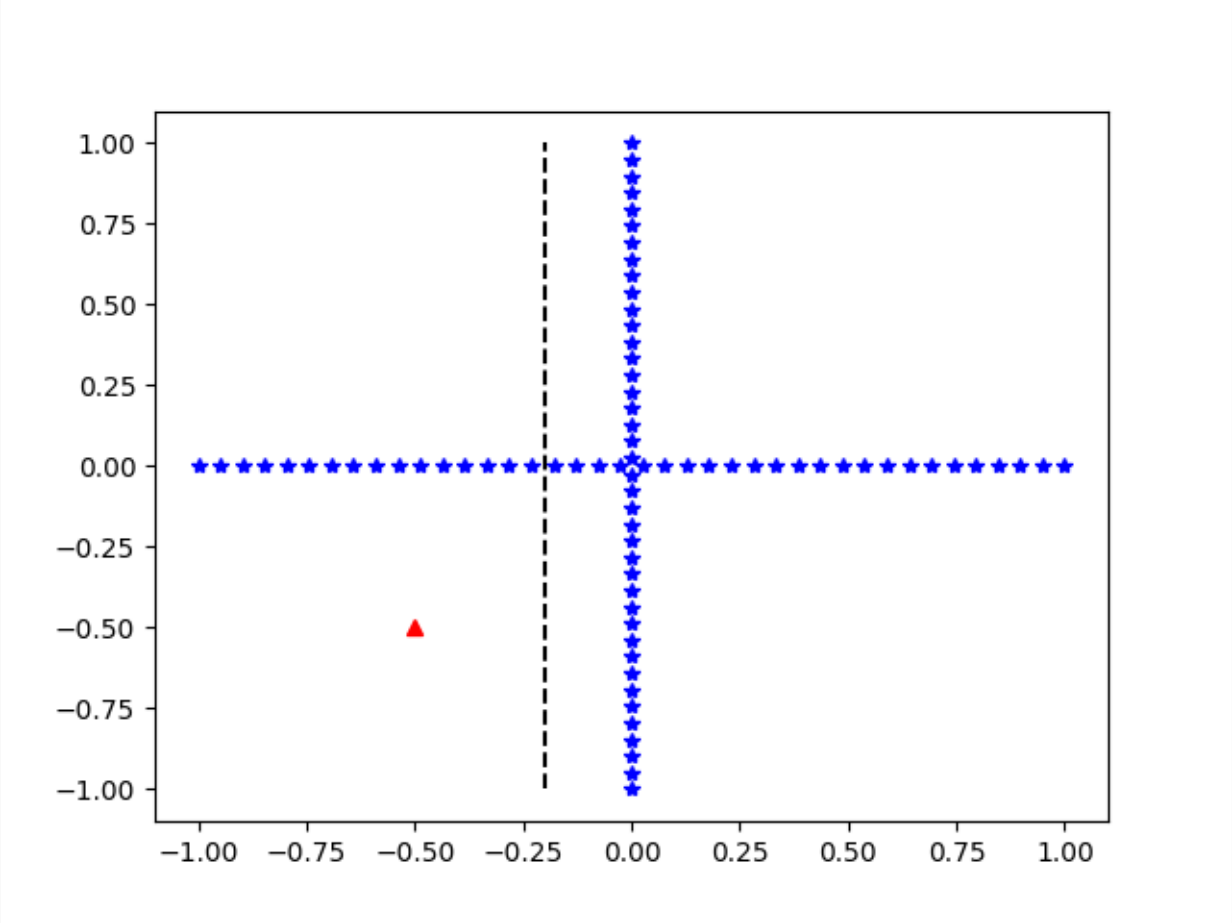
The IF algorithm **randomly partitions** the domain until all observations are **isolated**.

Then, an outlier score is assigned to each sample inversely proportional to the number of partitions, so that points that are isolated earlier get a higher score.

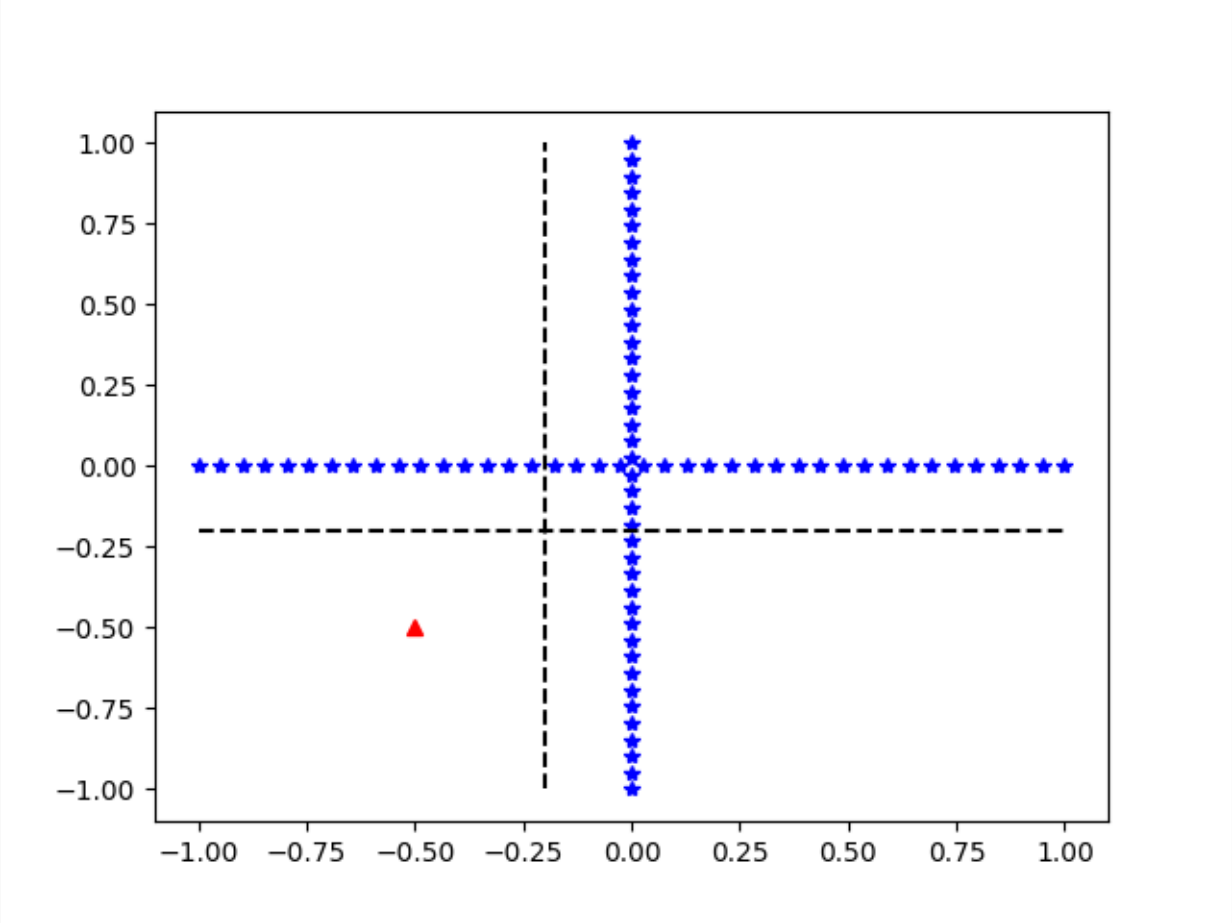
Low computational complexity.

Sharp decrease in performance in high dimensions.

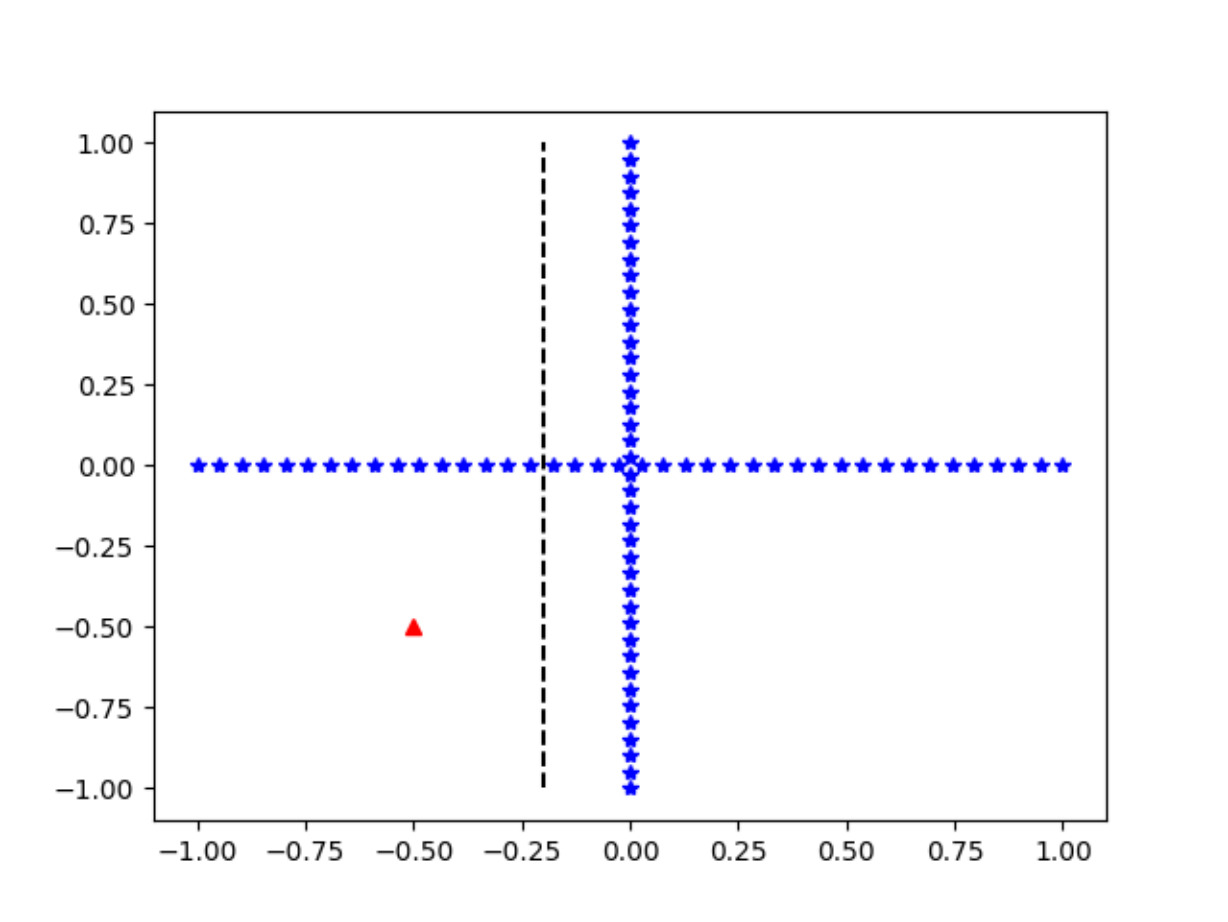
Example of isolation (outlier)



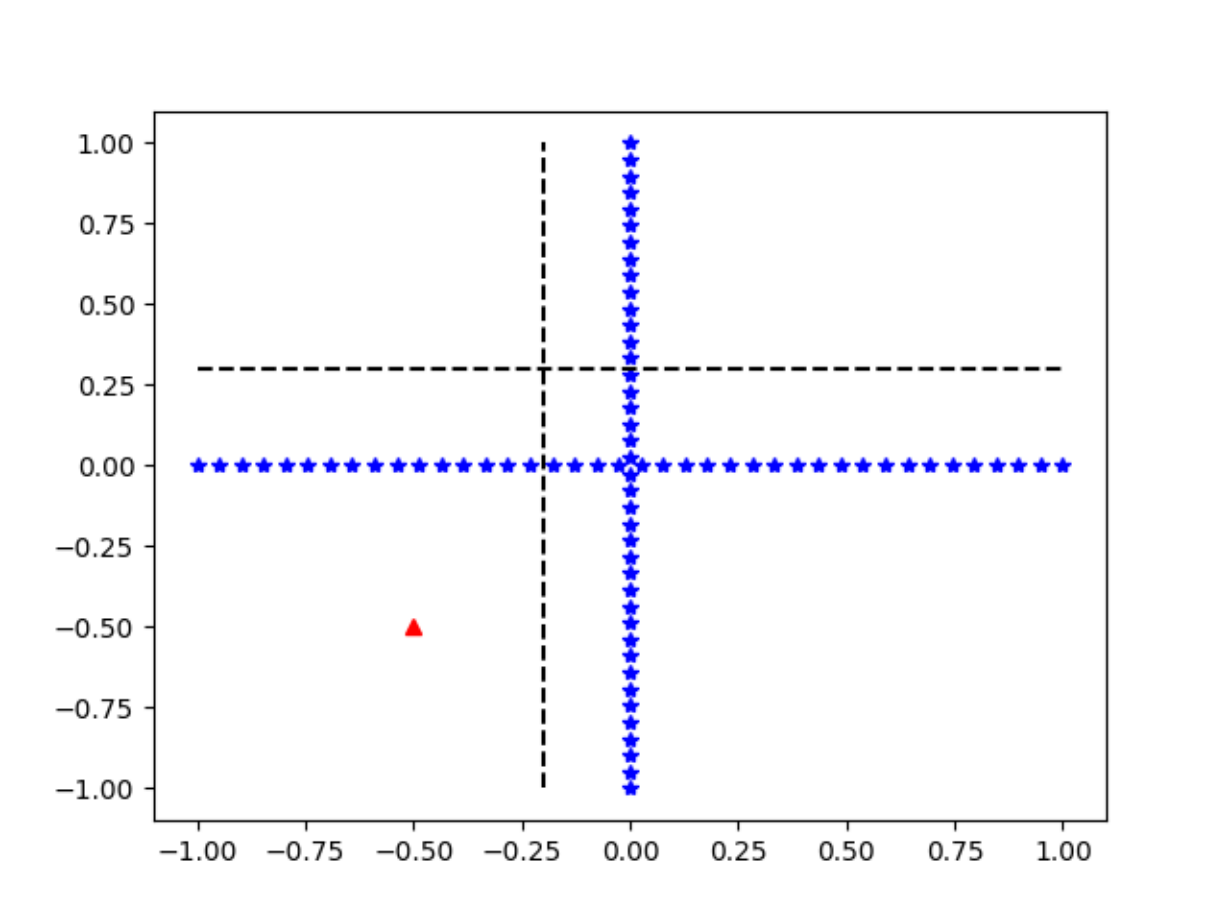
Example of isolation (outlier)



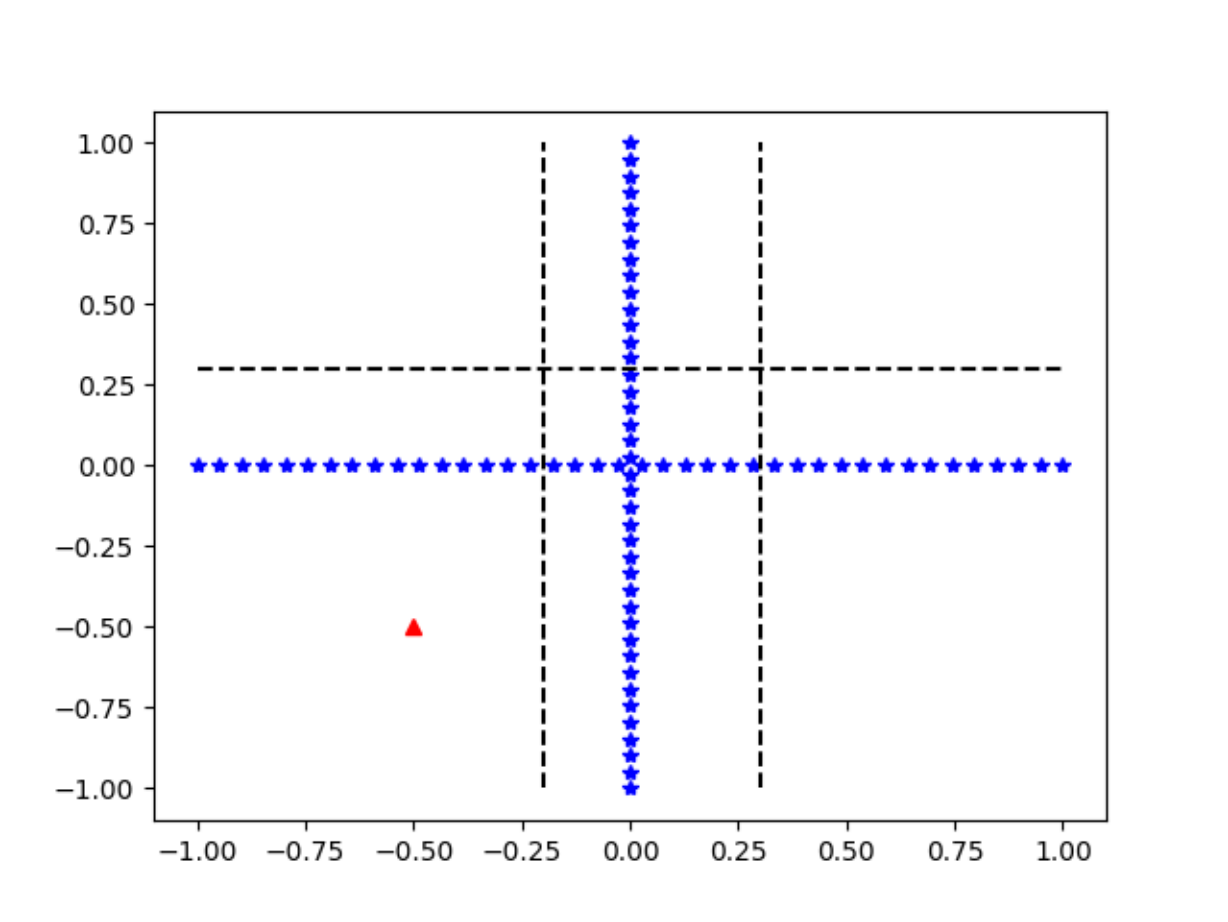
Example of isolation (inlier)



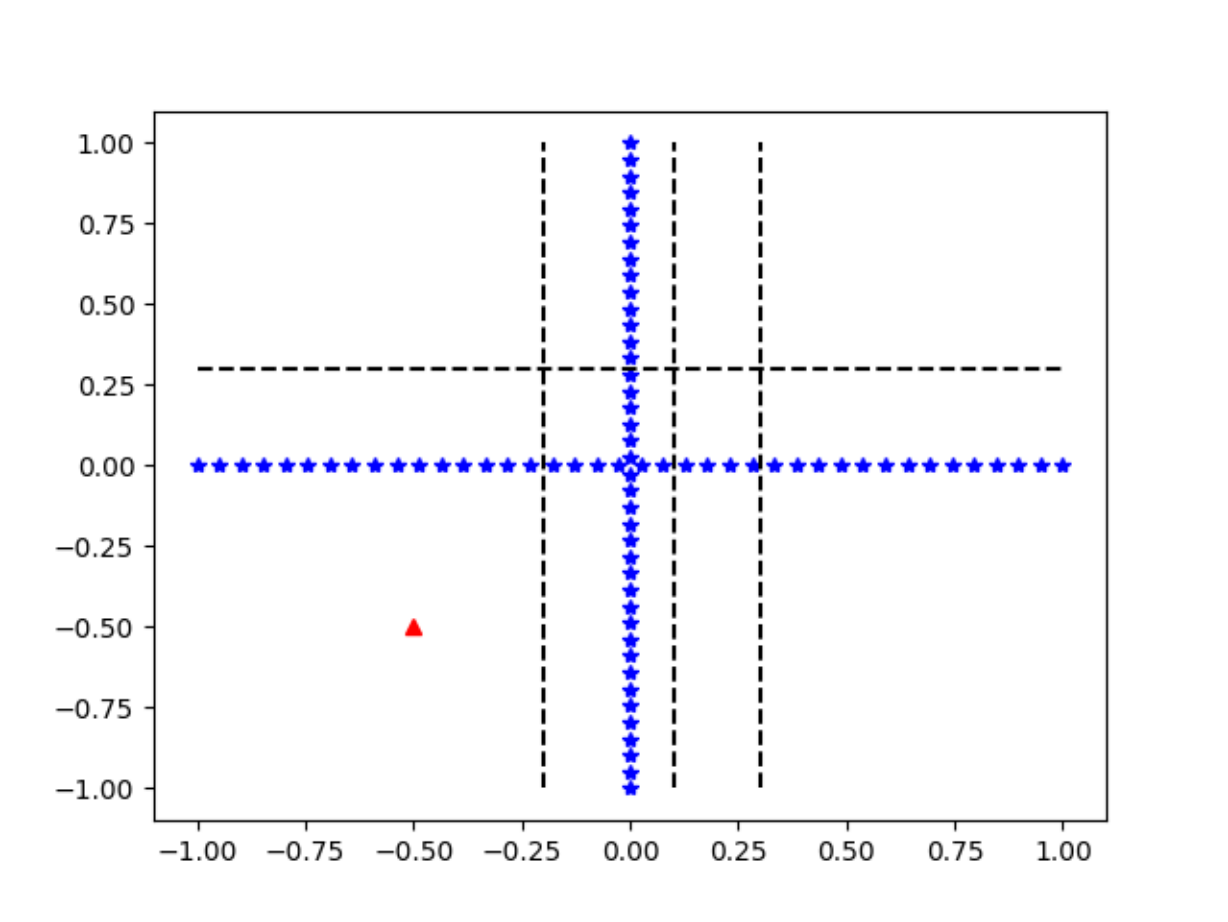
Example of isolation (inlier)



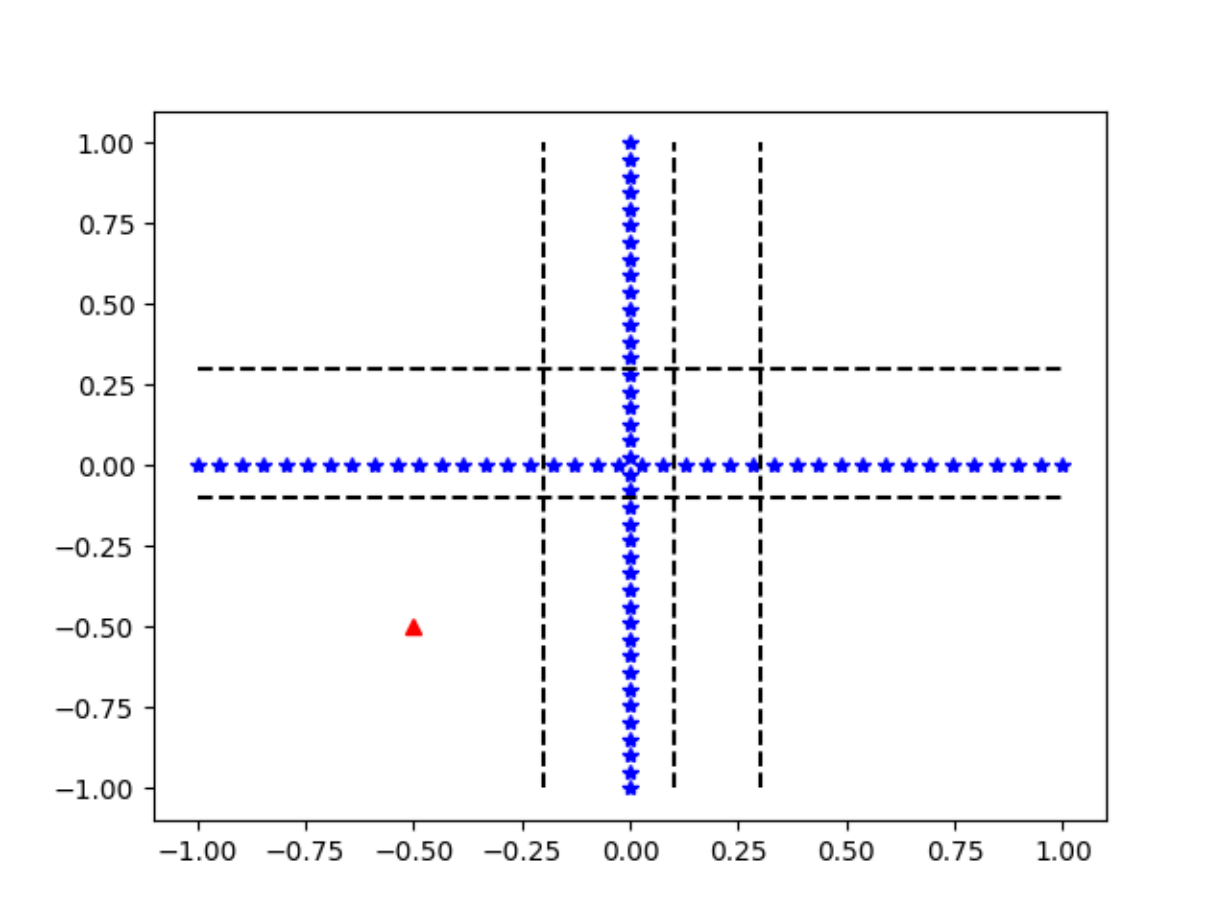
Example of isolation (inlier)



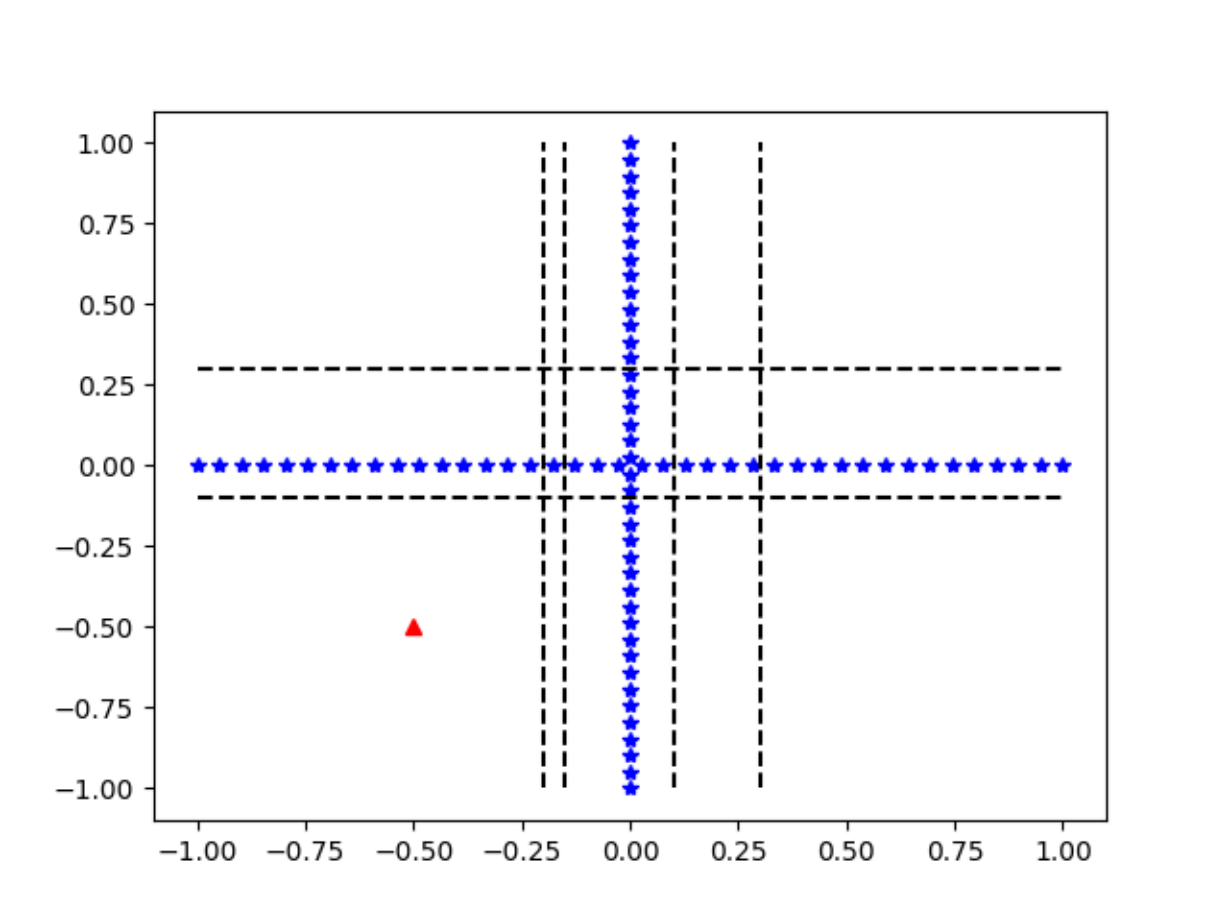
Example of isolation (inlier)



Example of isolation (inlier)



Example of isolation (inlier)



Can they detect hidden subspaces?

Both kNN and IF have low performances detecting anomalies in hidden subspaces when the number of features is high.

In the case of kNN, the concept of distant becomes unclear.

In the case of IF, the probability of choosing the right subspace decreases exponentially with the dimension of the subspace.

Loss of contrast

Lemma 1 (Aggarwal et al. 2001):

$$C_k \leq \lim_{d \to \infty} \mathbb{E}\left[\frac{Dmax_d^k - Dmin_d^k}{d^{1/k-1/2}}\right] \leq (n-1)C_k$$

Thus, the numerator **diverges** if we use the Manhattan distance, it is **bounded** for the Euclidean distance, and **zero** for any other distance metric.

No assumptions on the distribution of the data!

Choosing the right subspace

Lemma 2: Given a data set with d features, the probability of selecting a specific subspace of dimension k decays as $O(d^{-k})$.

$$p_k^M = \frac{k}{d} \sum_{i=1}^{M} \left(1 - \frac{k}{d}\right)^{i-1} p_{k-1}^{M-i}$$

Current proposal

We are currently working on a combination of the distance and isolation methods that aims to avoid the previous problems.

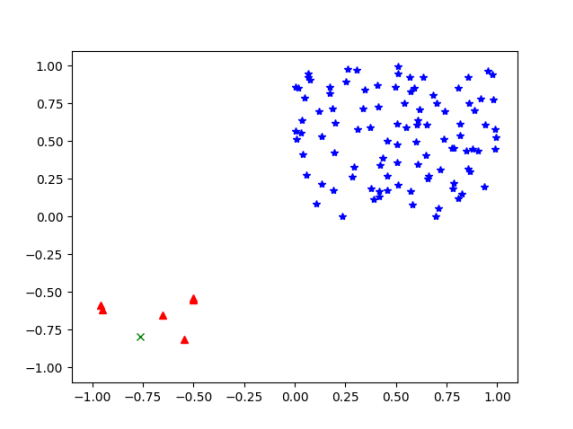
Given a certain point, we compute the distances to a random subsample.

This gives us a one-dimensional data set where the point of interest is in the left-fringe.

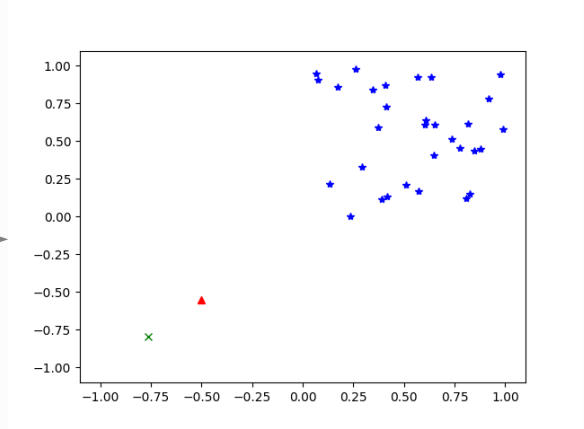
Then we compute the expected number of partitions to isolate said point and use that in the final score.

Example

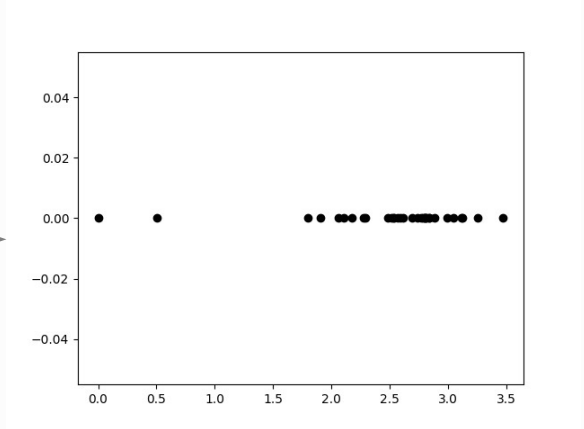
Select a point



Create a subsample



Compute distances



Isolation formula

We can benefit from the simplicity of the one-dimensional scenario to employ analytical formulas for isolation.

The expected number of partitions to isolate the left fringe in a data set of size n is given by:

$$\mathbb{E}[h] = \sum_{i=1}^{n-1} \frac{x_{i+1} - x_i}{x_{i+1} - x_1}$$

Fractional distance

We saw that the Manhattan distance is **more sensitive** in high dimensions than Euclidean or higher norms.

Could “distances” with **k<1** be even more sensitive?

In Aggarwal et al. (2001) they propose the use of **fractional distances**, which are even more sensitive than the Manhattan distance.

Test case

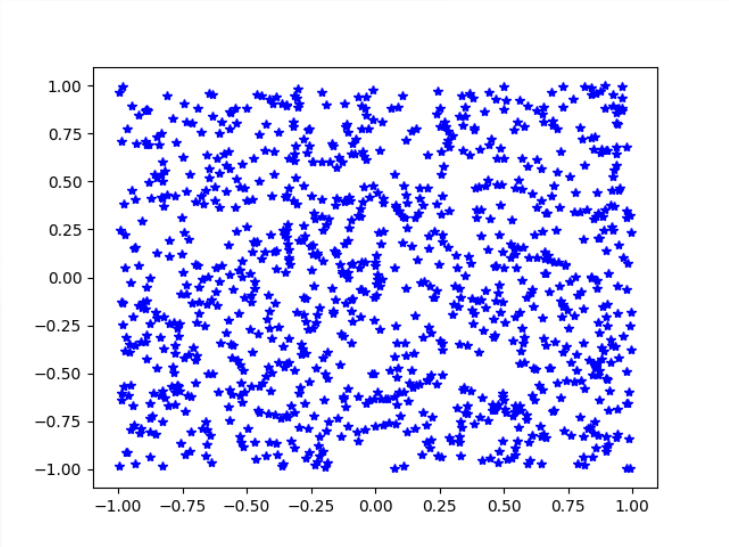
Data set of 1000 points with one clear outlier in a subspace of dimension 2.

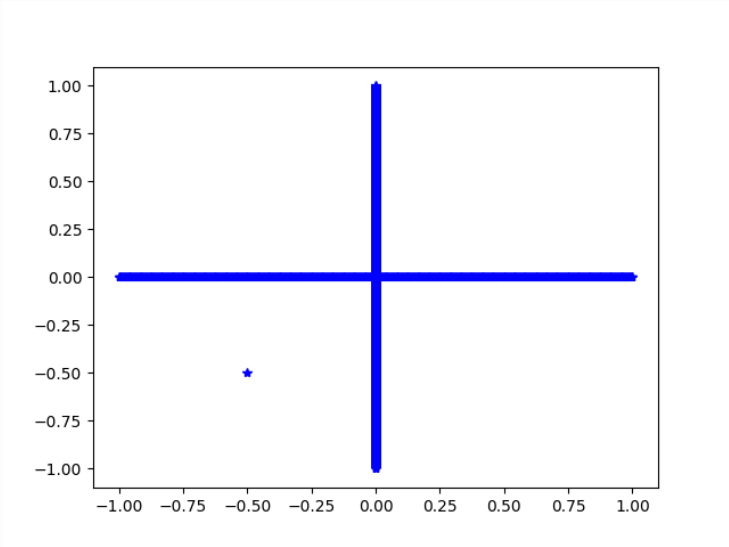
We add extra dimensions of i.i.d uniform random variables and test LOF, IF, and our proposed method for several levels of noise.

We also compare several distance metrics to study the loss of sensitivity in high dimensions.

Data set

All subspaces look “normal”, except for a hidden subspace of dimension 2.

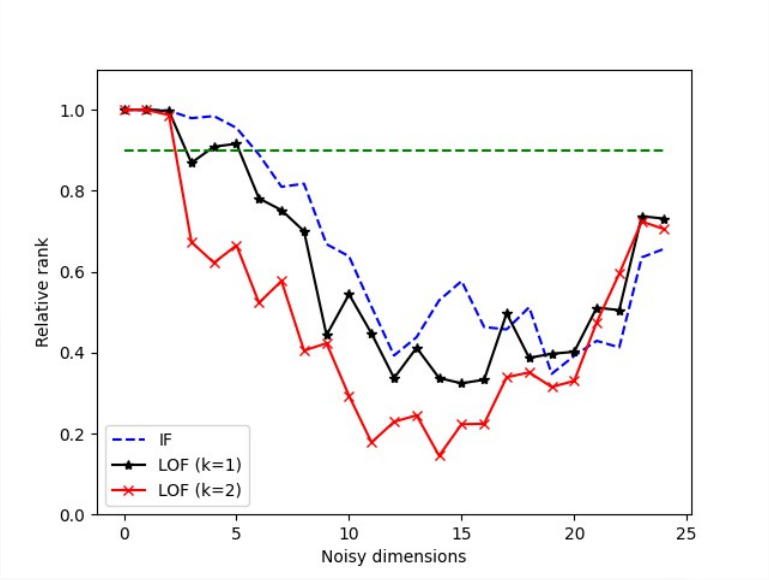


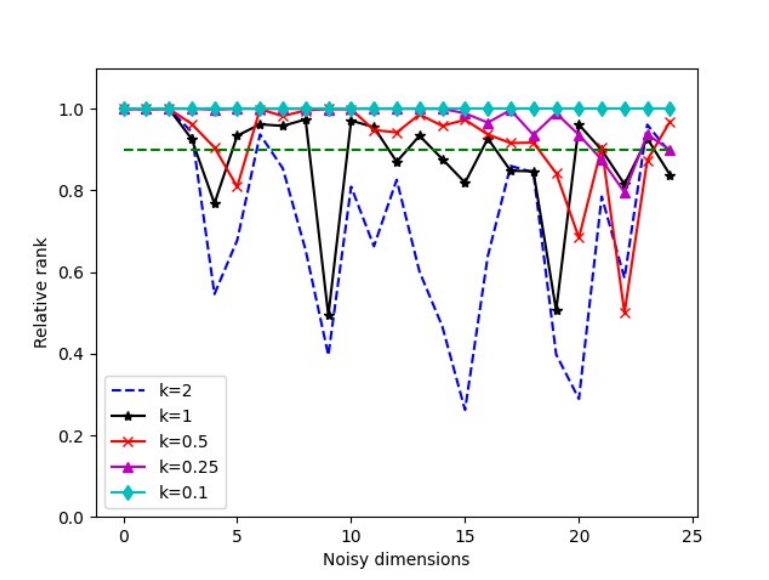


Results

Smaller values of k yield better results both using LOF and our proposed method.

IF and LOF do not detect the outlier for any d>5.





Pros and cons

+ It is able to detect small clusters of outliers.

+ It is less sensitive than IF to the dimensionality of the data.

- Loss of sensitivity in very high dimensions.

- Higher computational complexity than IF.

Further research

- Inclusion of categorical features.
- Further testing of distance metrics and subspace search algorithms.
- Explainability methods.
- Inclusion of time series data.

References

- C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In J. Van de Bussche and V. Vianu, editors, Lecture Note in Computer Science, volume 1973. Springer, Berlin, Heidelberg, 2001. doi:10.1007/3-540-44503-X_27
- F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation-Based Anomaly Detection. ACM Transactions on Knowledge Discovery from Data, 6(1):1–39, 2012. doi:10.1145/2133360.2133363.