Outlier Detection Methods

Paul van Leeuwen

5 December 2019

Introduction

Traditional Methods

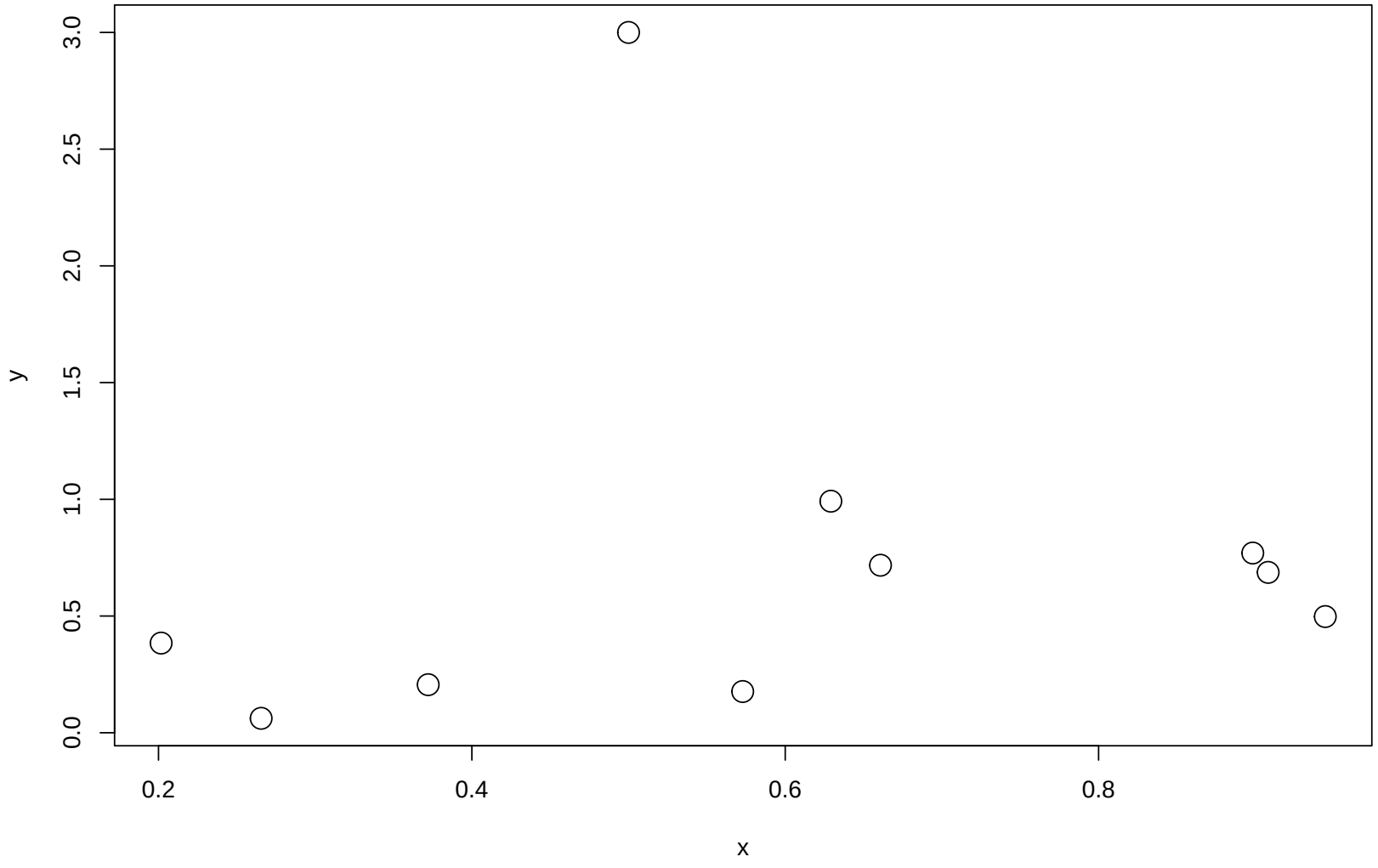
- (Hawkins-Outlier, 1980) 'An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.'
- Traditional outlier detection methods can be categorised into the following approaches:
  - distribution-based: easy to visualise but a multivariate probability distribution needs to be assigned to all variables, which is unknown in our case;
  - depth-based: outliers are assumed to be located at the boundaries of the data and computational demanding for four or more dimensions, which is applicable to our case;
  - clustering: methods are optimised to cluster the data, not to detect outliers;
  - distance-based: problematic when we have sparse and dense data regions, which could easily be the case for high levels of the LOB.

A Novel Approach

- M. Breunig, et al. introduced a new approach: Local Outlier Factor (LOF).
  - This is a density-based approach driven by the data.
  - Data points that are distant *relative* to eachother are considered to be more outlying.
  - Issues above are more or less solved, although we still need to properly define the parameters.
  - In addition, the variables need to be continuous and outliers in low density regions are still hard to detect.
- This inspired variants, worth to be investigated:
  - Connectivity-based Outlier Factor (COF) by Tang et al. 2002;
  - Influenced Outlierness (INFLO) by Jin et al. 2006;
  - Local Outlier Correlation Integral (LOCI) by Papadimitriou et al. 2003;
  - ...
- A great overview of these methods are given in https://archive.siam.org/meetings/sdm10/tutorial3.pdf.

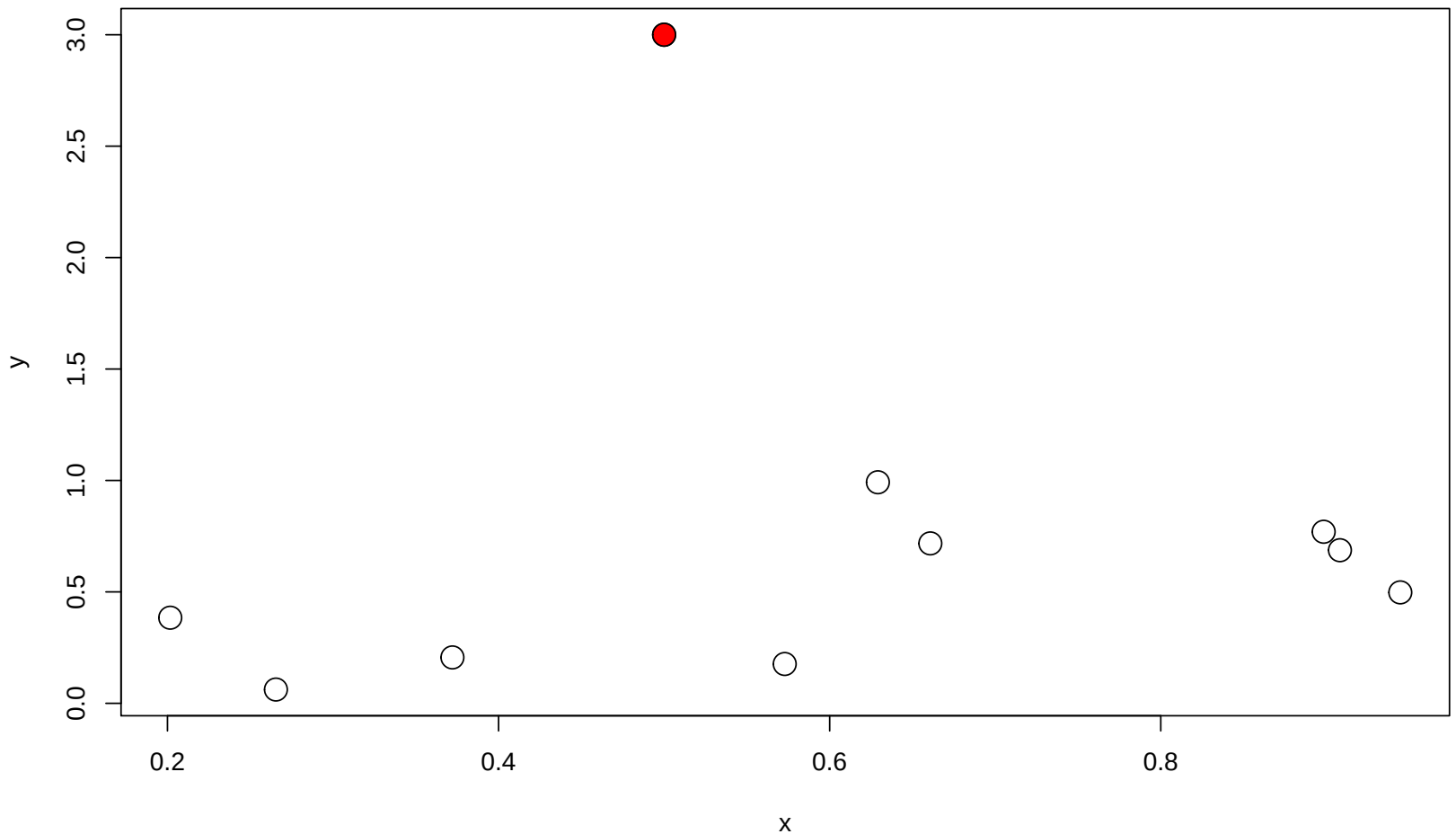
How Does LOF Work?

How Does LOF Work?



How Does LOF Work?

- Without any knowledge of the probability distribution we could have assigned to the data, the point (0.5, 3) is considered to be an outlier.



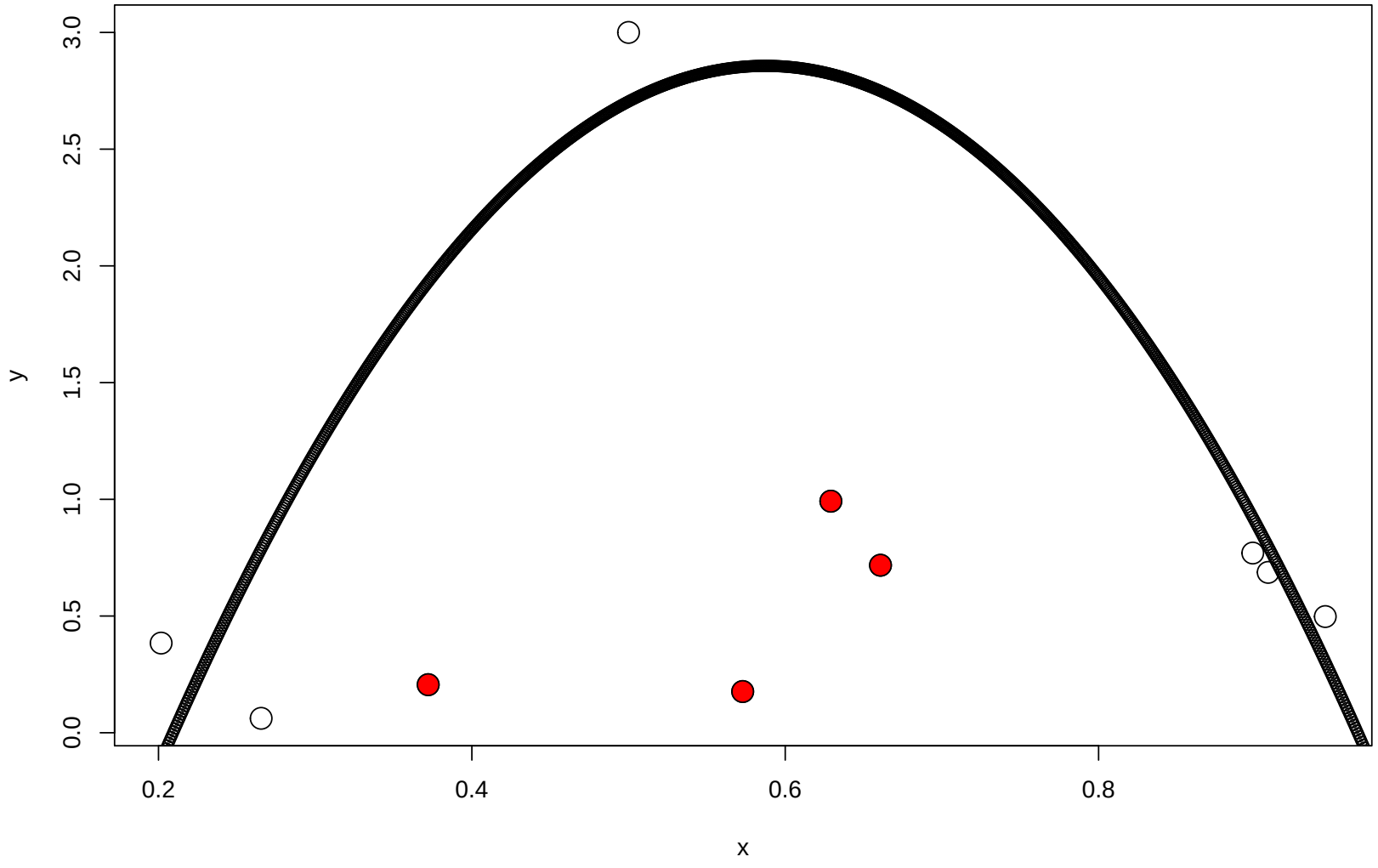
How Does LOF Work?

- However, suppose that *a priori* we know that the data points (x_i, y_i) for $i = 1, \ldots, 10$ follow the pattern

$$y_i = -4.04 + 23.5x_i - 20x_i^2 + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, 0.933)$$

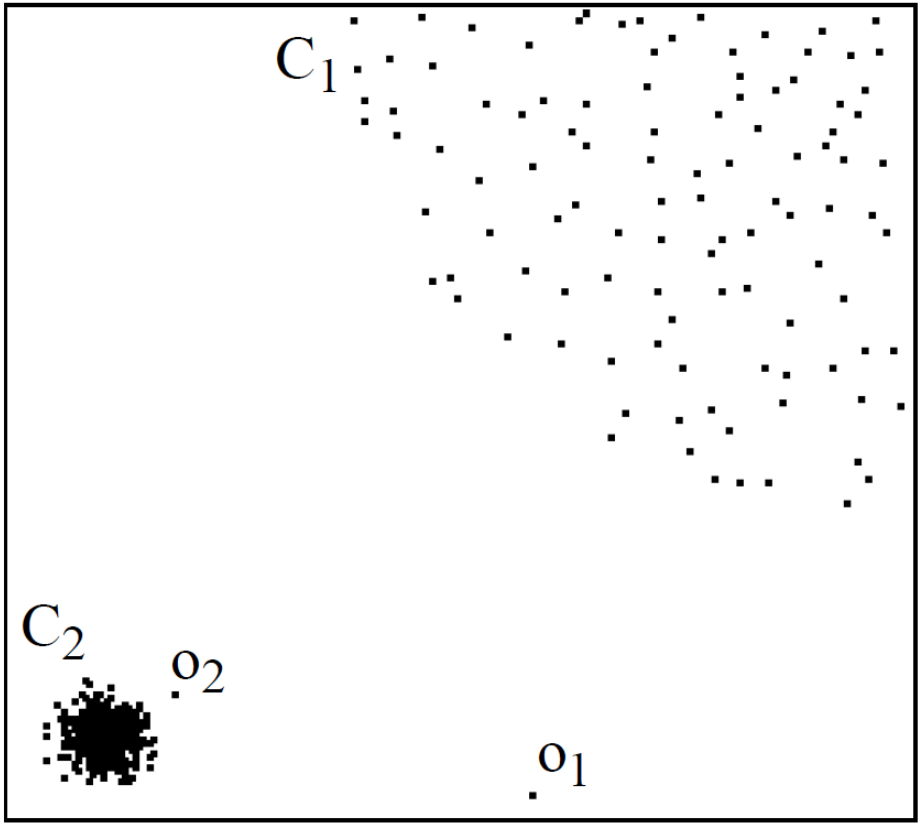
  - A second-order polynomial is fitted on the data points leaving the ones out that meet the conditions $0.3 < x_i < 0.8$ and $y_i < 1.5$.
- Then the point considered to be an outlier before is not an outlier anymore, but the points that are left out are!

How Does LOF Work?



How Does LOF Work?

- However, in our case we do not have that level of knowledge of the data-generating process of y_i.
- Alternatively, make use of the relative densities.
- The figure below is retrieved from M. Breunig, et al.



How Does LOF Work?

- The traditional methods have a hard time dealing with different densities.
  - For example, the algorithms from the distance-based approach cannot identify o_1 as an outlier while the points in the cluster C_2 are not.
- Make use of the Eucledian distance.
  - Is standardisation necessary?
- For each data point investigate how dense the neighbourhood is for each of its k neighbours.
- First, calculate the reachability distance of all data points.
- Second, calculate the local reachability of each data point.
  - Calculate the inverse of the average of reachability distances of its k nearest neighbours.
- Finally, the LOF of a data point is the local reachability of its k nearest neighbours *relative* to the local reachability of that data point.

The LOF Algorithm

- $\mathit{reach\text{-}dist}_k(p, o) = \max\{k\text{-distance}(o), \mathit{dist}(o, p)\}$
- $kNN(p)$ is in practice the set k nearest neighbours.
- $\mathit{lrd}_k(p) = \left(\frac{\sum_{o \in kNN(p)} \mathit{reach\text{-}dist}_k(p, o)}{|kNN(p)|} \right)^{-1}$
- $LOF_k(p) = \frac{\sum_{o \in kNN(p)} \frac{\mathit{lrd}_k(o)}{\mathit{lrd}_k(p)}}{|kNN(p)|}$

How Does LOF Work?

- A LOF-value around (way above) one is considered to be an inlier (outlier).
  - In the figure retrieved from M. Breunig, et al. all data points of the clusters C_1 and C_2 are inliers while the data points o_1 and o_2 have a value clearly more than one.
- However, the choice for the number of nearest neighbours k remains ambiguous.
  - M. Breunig, et al. provide some heuristics on the minimum and maximum values of k, but this remains vague and additional information on the data-generating process is required.
- Another issue is that, even is k chosen appropriately, some clusters are not properly identified. Or what about outlying clusters?
- Finally, how do we deal with categorical values?

An Alternative to LOF

LOCI

- To deal with the arbitrary choice of number of nearest neighbours k the Local Outlier Correlation Integral (LOCI) method is introduced.
  - This approach resembles the LOF-method.
  - Differences arise as the neighbourhood is much more continuous, instead of discrete and rather arbitrary.
  - Although some parameters need to be chosen beforehand, k is automatically dealt with.

LOCI

- Questions to be answered for LOCI:
  - Chebyshevs' inequality

$$\mathbb{P}[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}, \quad k > 1$$

    is used for a random variable X with expected value μ and standard deviation σ. But the method uses the sample standard deviation while Chebyshevs' inequality uses the population standard deviation. And there are more efficient alternatives available, such as the upper probability bound provided by Saw et al. (1984).
  - What is influence of the parameters α and k? And why are they set at $\alpha = 0.5$ and $k = 3$?
  - Is 20 as chosen in the paper the appropriate minimum number of neighbours to start with? Is it much affected by the choice of the population probability function?
  - Example outliers in the paper are hard to reproduce.