# A review of outliers detection

Titouan Vayer

May 17, 2017

# Sommaire

Les travaux suivants sont basés essentiellement sur le papier de [Victoria J.Hodge : A Survey of Outlier Detection Methodologies]

# Qu'est-ce qu'un outlier ?

Deux définitions possibles :

- (Grubbs, 1969) Un outlier est une observation qui semble dévier fortement par rapport aux autres observations du sample dans lequel il se situe
- (Barnett & Lewis, 1994) Une observation (ou un ensemble d'observations) est un outlier s'il il apparaît comme étant contradictoire avec le reste des données.

# Les trois grandes approches pour la détection d'outlier

## Comment mettre en place une stratégie de détection ?

- **Unsupervised clustering** : Déterminer l'outlier sans à priori sur les données. Suppose d'avoir un dataset suffisamment fourni (type I)
- **Supervised classification** : Modéliser la normalité et la normalité. Requiert d'avoir des données labélisées. type(II)
- **Semi-supervised classification** : Modéliser seulement la normalité ou alors l'anormalité. Utilisable pour des données figées et non figées, apprend et s'améliore au fur et à mesure que les données arrivent. (type III)

# Comment traiter un outlier ?

On a deux approches possibles pour traiter un outlier :

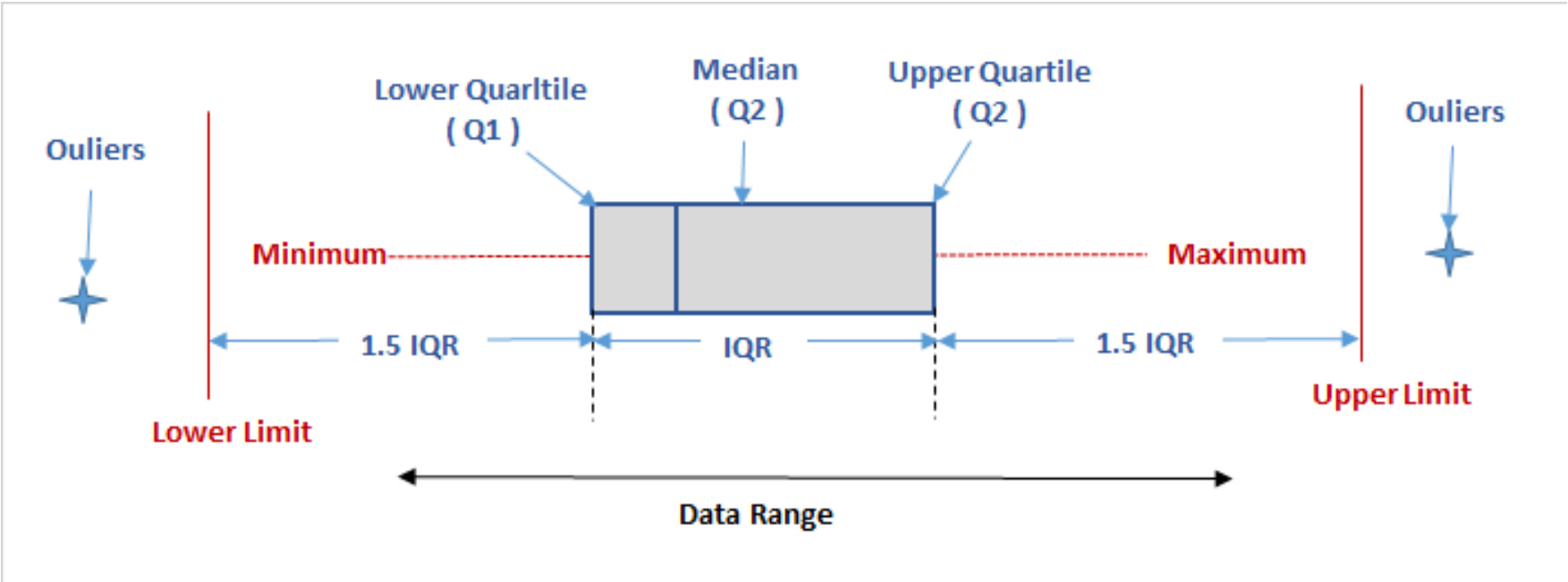
- **Le diagnostique** : on analyse les outliers et on les enlève ou pas
- **L'accomodation** : on garde les outliers quoiqu'il se passe

# Introduction

Les approches statistiques sont les algorithmes les plus anciens pour détecter les outliers.
Ils sont généralement bien dimensionnés pour des données quantitatives et s'intéressent à la distribution des données.

# Boxplot

C'est une technique de détection **non supervisée**. L'idée est de produire une représentation graphique qui permet de juger si des points sont outliers ou non.




Laurikkala et al. suggère une distance (heuristique) de 1.5 inter-quartile range beyond entre la limite haute et basse pour détecter les outliers.

# ESD test ou test de Grubb

C'est une technique de détection **non supervisée**.
The Extreme Studentized Deviate ou ESD test (Rosner 1983) est utilisé pour détecter un ou plusieurs outliers pour des données univariées qui suivent approximativement une **distribution normale** . Il teste les hypothèses suivantes :

- H0: Il n'y a pas d'outlier
- Ha: Il y a au moins un outlier

Le test statistique est le suivant :

$$G = \frac{\max_{i=1,\ldots,N} \left| Y_i - \bar{Y} \right|}{s}$$

Le test de Grubb correspond à la plus grande déviation par rapport à la moyenne par rapport à la variance.

# ESD test ou test de Grubb

Il peut être aussi défini en tant que "one-side test", pour tester si la valeur minimale est un outlier avec

$$G = \frac{\bar{Y} - Y_{\text{min}}}{s}$$

ou alors

$$G = \frac{Y_{\text{max}} - \bar{Y}}{s}$$

Pour le cas général, l'hypothèse nulle est rejetté pour un degré $\alpha$ si

$$G \geq \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2} + t^2_{\alpha/(2N),N-2}}$$

avec $t^2_{\frac{\alpha}{2N},N-2}$ représente la borne supérieure de la t-distribution à $N-2$ degrés de liberté.

# Mahalanobis distance

C'est une technique de détection **non supervisée**.

Contrairement à la méthode précédente elle permet de prendre en compte l'aspect multivarié d'un dataset. La méthode se base sur le principe suivant : si $X \sim \mathbb{N}(\mu, \Sigma)$ alors $D^2(X, \mu) \sim \chi^2_p$ où $D^2(X, \mu)$ est la distance de Mahalanobis définie par :

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}$$

avec $\mu$ et $S$ la moyenne et la matrice de variance-covariance du dataset.

On peut donc avoir un intervalle de confiance de tel sorte que :

$$\mathbb{P}[D^2(X, \mu) \leq \chi^2_{p,1-\alpha}] \leq 1 - \alpha$$

# Implémentation

La détection par la distance de Mahanalobis a été implémentée sur le github suivant [Titouan Vayer]
`https://github.com/bigtdu53/outlierdetection`

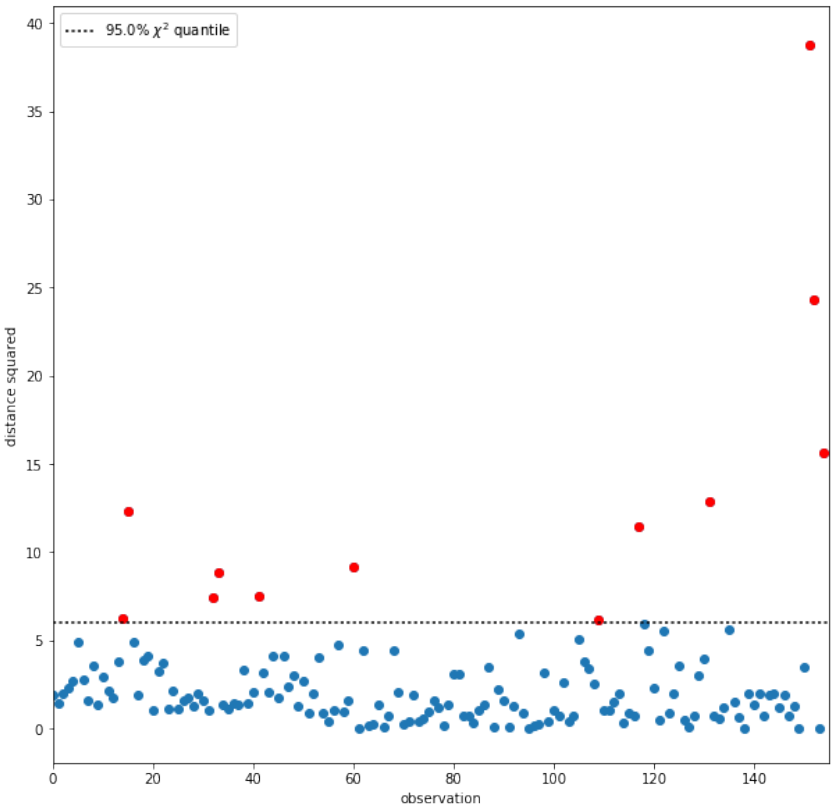
Les données utilisées ont été :

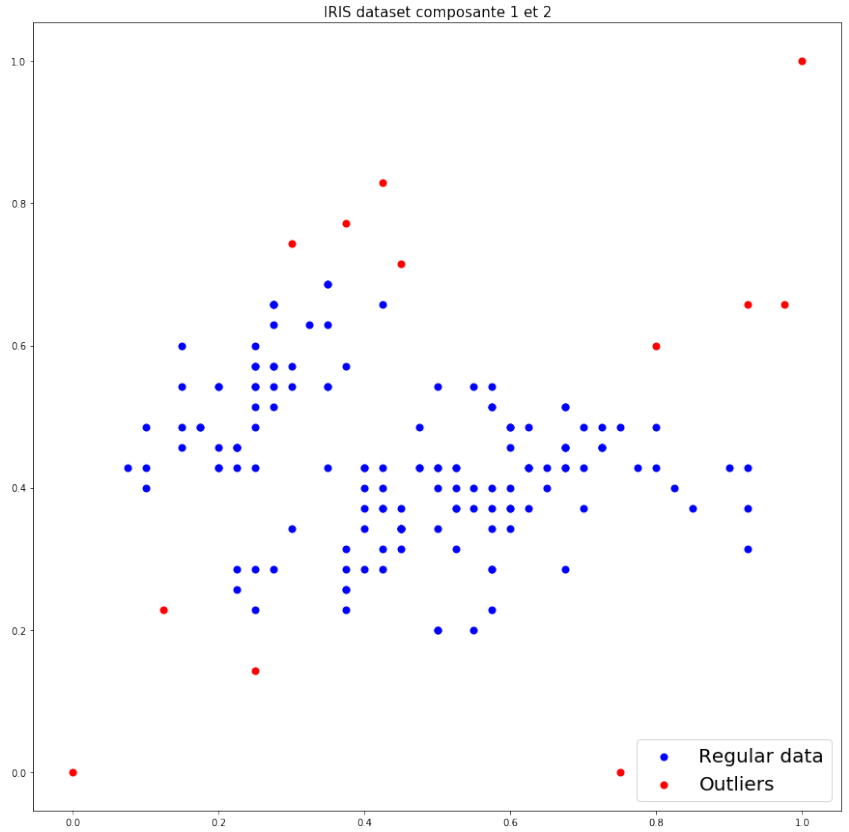
- Iris Dataset
- dtmcross201606.csv

L'inconvénient présente deux incovénients :

- La normalité des données
- L'inversion de la matrice très coûteuse en grande dimension

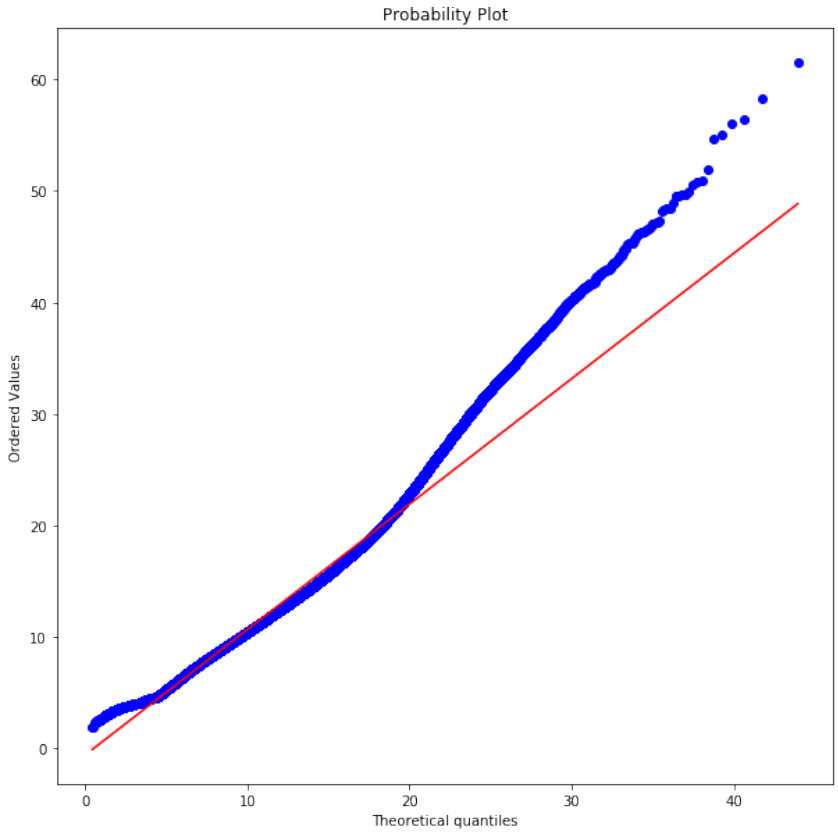
# Outlier detection using Mahanalobis distance on Iris Dataset

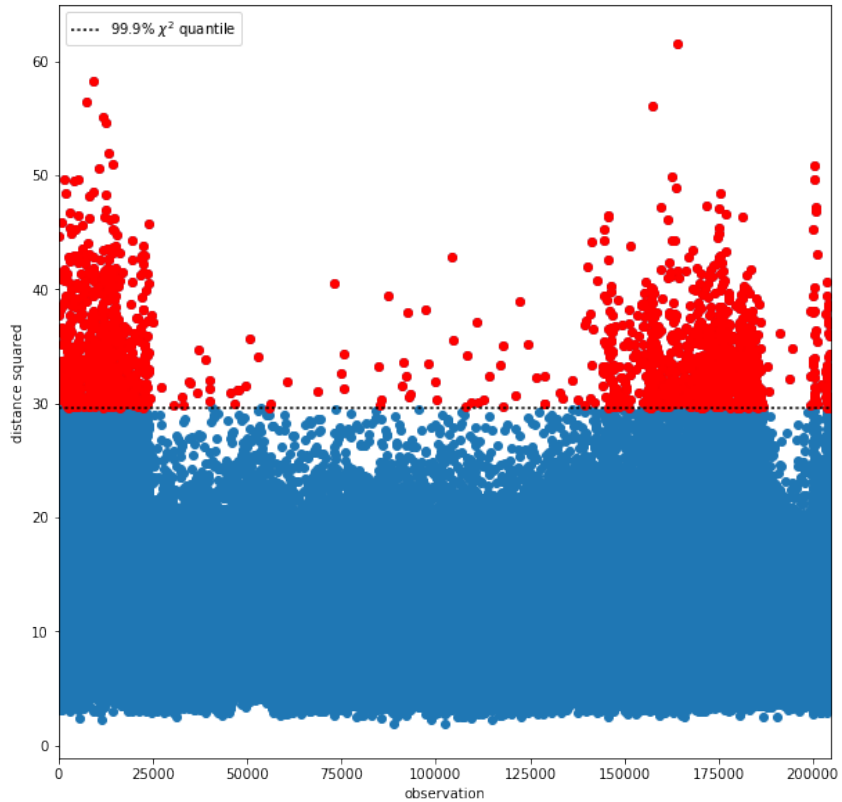


# Outlier detection using Mahanalobis distance on dtmcross

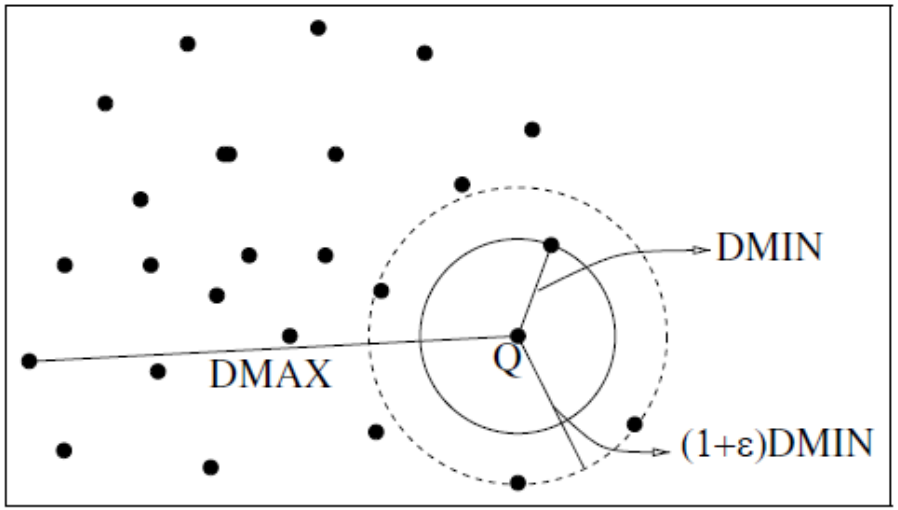
Dans le cas de dtmcross on a une très grand dataset (environ 1000 variables). Aussi pour pouvoir appliquer la méthode on passe d'abord par une étape (PCA) de réduction de dimension.



# Limitation : Malédiction de la dimension

Plus la dimension augmente, plus les points sont regroupés dans un volume plus grand et qui devient moins dense.([Kevin Beyer])




Les modèles statistiques utilisent différentes approches pour s'affranchir du probèle de la malédiction de la dimension, toutes ces techniques entrainent un grand coup de processing. Une autre alternative est de réduire la dimension de l'espace.

# Le théorème de Beyer

On considère $X_1, ..., X_d \in \mathbb{R}^n$ de telle sorte que
$\forall m \in [1, ..., d], X_{m,1}, ..., X_{m,n}$ sont $n$ points indépendants tirés selon une distribution $\mathcal{F}_m$.
On choisit un point $Q_m$ indépendamment des $X_{m,i}$. On note :
$d_{\min}(m) = \min\{d(X_{m,i}, Q_m) | 1 \leq i \leq n\}$ et
$d_{\max}(m) = \max\{d(X_{m,i}, Q_m) | 1 \leq i \leq n\}$

## Theorem

$$\text{Si } \lim_{d \to +\infty} var\left(\frac{d(X_{m,1}, Q_m)^p}{\mathbb{E}[d(X_{m,1}, Q_m)^p]}\right) = 0 \text{ alors}$$

$$\forall \epsilon \geq 0, \mathbb{P}(d_{\max}(m) \leq (1+\epsilon) d_{\min}(m))) = 1$$

Ce théorème est valable par exemple lorsque les données sont i.i.d dans chaque dimension, que les moments sont finis et que les query point sont choisis indépendamment vis à vis des données

# Proximity based techniques

Proximity-based techniques are simple to implement and make no prior assumptions about the data distribution model. They are suitable for both type 1 and type 2 outlier detection. However, they suffer exponential computational growth as they are founded on the calculation of the distances between all records. The computational complexity is directly proportional to both the dimensionality of the data m and the number of records n

# K-Means

K-means initially chooses random cluster prototypes according to a user-defined selection process, the input data is applied iteratively and the algorithm identifies the best matching cluster and updates the cluster centre to reflect the new exemplar and minimise the sum-of-squares clustering function given by equation :

$$\sum_{j=1}^{K} \sum_{n \in S_j} ||x^n - \mu_j||^2$$

Dragon use an adapted similarity metric that incorporates the word count from the news story, the distance to the clusters and the effect the insertion has on the prototype vector for the cluster. After training, each cluster has a radius which is the distance between the prototype and the most distant point lying in the cluster. This radius defines the bounds of normality and is local to each cluster rather than the global distance settings used in many approaches such as (Knorr and Ng, 1998), (Ramaswamy et al., 2000) and (Byers and Raftery, 1998) k-NN

# References


Victoria J.Hodge (2017)

A Survey of Outlier Detection Methodologies

http://book.itep.ru/depository/security/anomaly/Hodge+Austin_OutlierDetection_AIRE381.pdf

Titouan Vayer Github (2017)

GitHub repository

https://github.com/bigtdu53/outlierdetection

When is Nearest Neighbor Meaningful ? (1998)

When is Nearest Neighbor Meaningful ?

https://members.loria.fr/MOBerger/Enseignement/Master2/Exposes/beyer.pdf