Détection des anomalies

Local Outlier Factor

Ricco Rakotomalala

Université Lumière Lyon 2

Plan

1. Problématique
2. Local Outlier Factor
3. Un exemple
4. Conclusion
5. Références

Quelques définitions - Problématique

DÉTECTION DES ANOMALIES

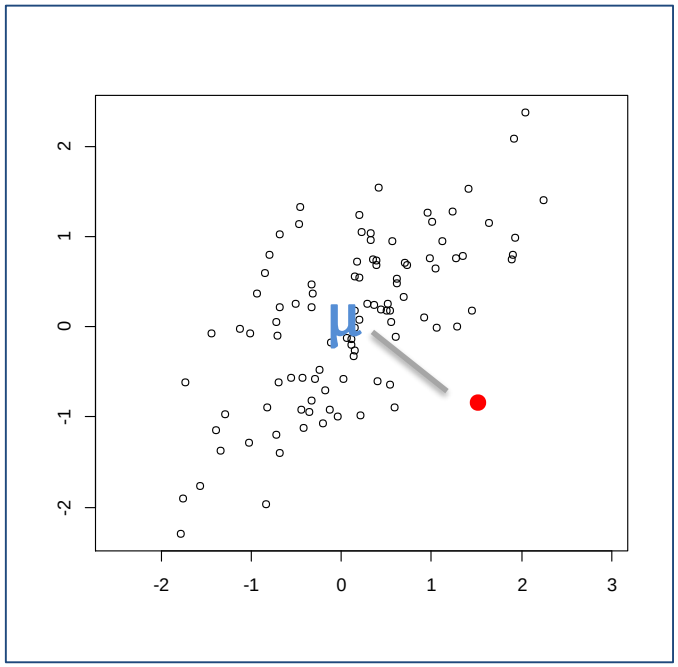
2 formes d'anomalies à identifier dans les données

Détection des points atypiques (outlier detection) : un ou des points s'écartent significativement des autres dans une base de données. Ils sont épars et localisés dans une zone peu dense des données (s'ils forment un groupe compact, on ne peut pas vraiment parler d'anomalies)

Détection des nouveautés (novelty detection) : on situe un individu supplémentaire par rapport à un échantillon de référence (considéré « propre »), on cherche à savoir s'il peut y être associé ou s'il s'en écarte significativement

Approche simple – Distance de Mahalanobis

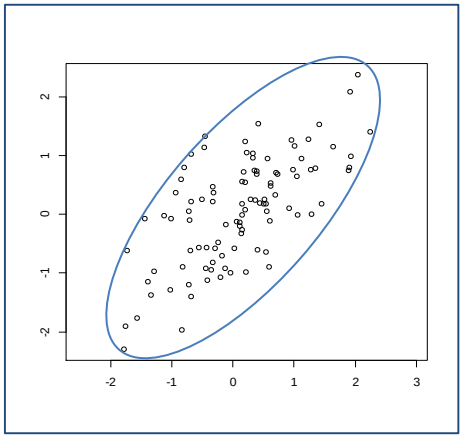
Pour chaque point, calculer la distance par rapport au barycentre (μ) – qui sert de référence – en tenant compte de la forme du nuage de points (via la covariance Σ).



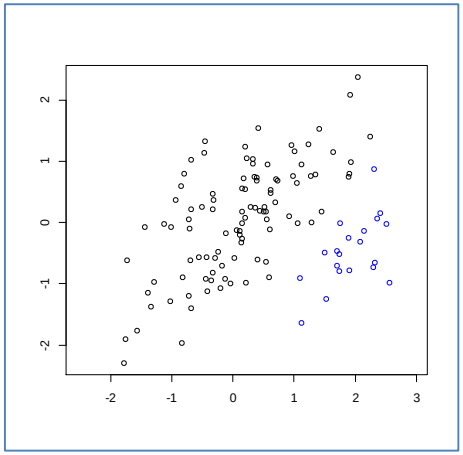
$$d^2_{(\mu,\Sigma)}(x_i) = (x_i - \mu)'\Sigma^{-1}(x_i - \mu)$$

Le point rouge n'est pas atypique sur les deux axes pris individuellement, mais l'est par rapport à la forme du nuage de points

Problème distance de Mahalanobis

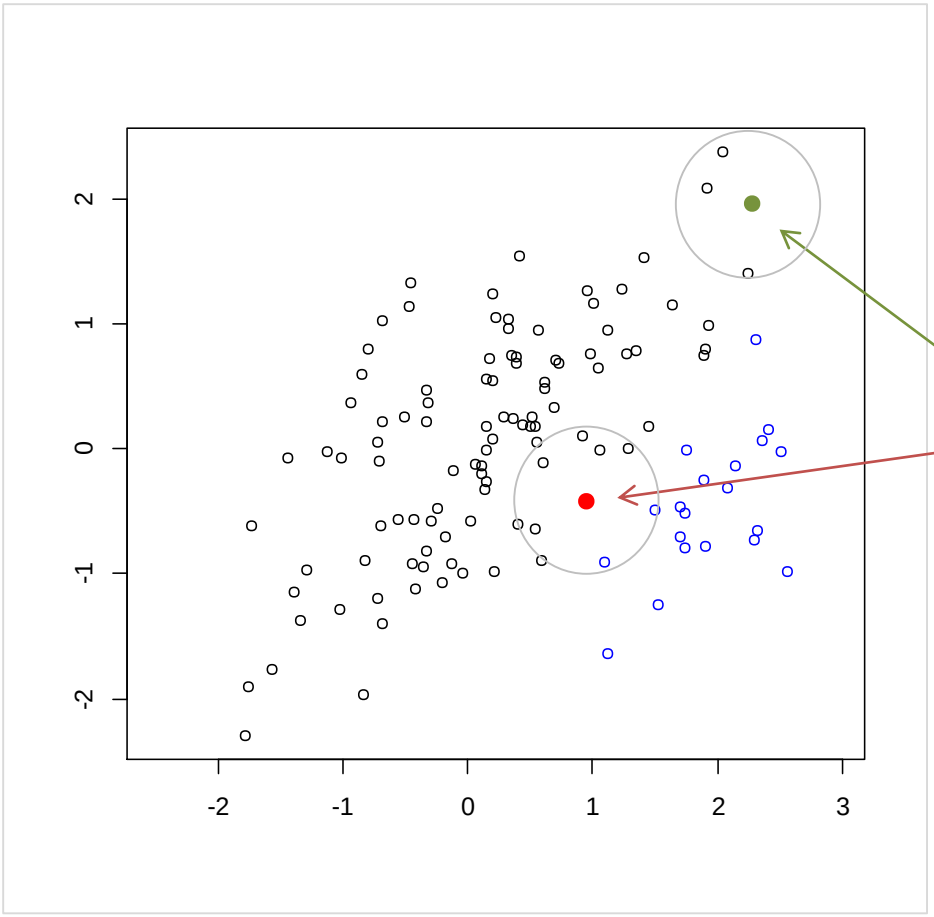


Les calculs de μ et Σ peuvent être affectés par les points atypiques (Remarque : des solutions robustes existent… ex. « Minimum Covariance Determinant estimator », Rousseuw, 1984)



Les données peuvent être non-gaussiens, ou clustérisées, le barycentre global ne veut plus rien dire.

Identification locale des points atypiques



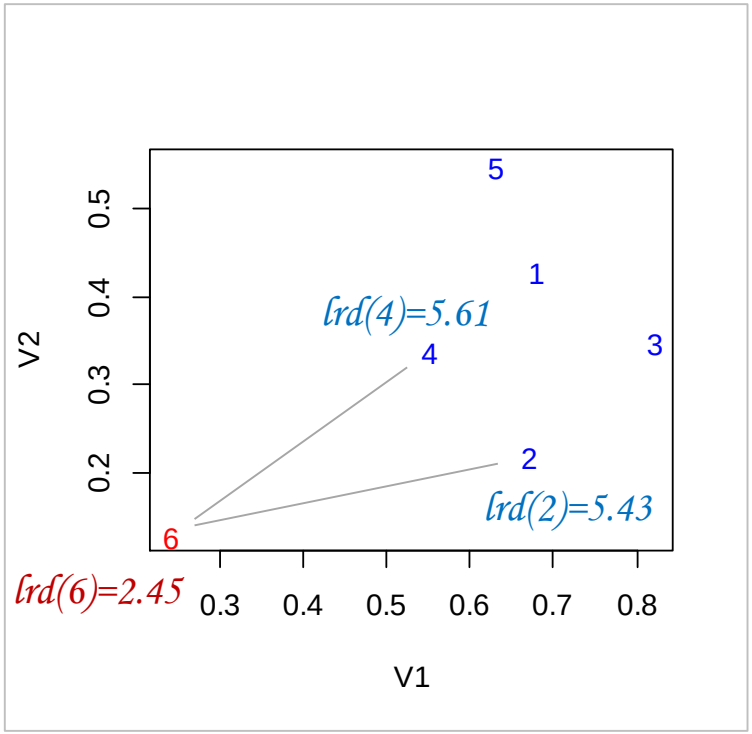
Dans une zone à forte densité, un point qui s'écarte des autres (de ses voisins immédiats) devrait plus interroger que lorsqu'il se situe dans une zone moins dense.

Calcul de densité locale basée sur les k-plus proches voisins

LOCAL OUTLIER FACTOR

Principe du Local Outlier Factor (LOF)

Comparer la densité locale d'un point avec celles de ses k (paramètre) plus proches voisins : si elle est inférieure, suspicion de point atypique



Exemple (k = 2)

$$\begin{cases} lrd(6) \ll lrd(4) \\ lrd(6) \ll lrd(2) \end{cases}$$

Le point n°6 est potentiellement atypique

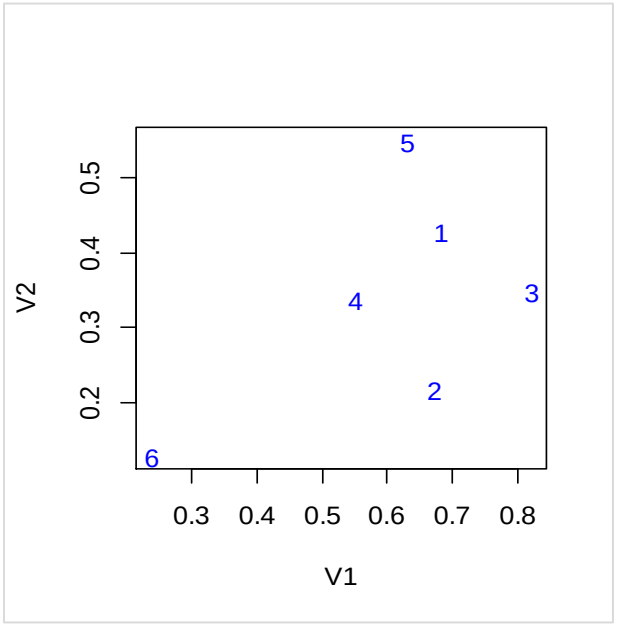
$$lof(6) = 2.2459 \ldots \gg 1$$

Valeur de référence

lrd (local reachability density) : mesure de densité locale d'un point (↑ fortement entouré, densité élevée ; ↓ faiblement entouré, densité faible)

Quelques définitions et calculs

$$d^2(i,i') = \sum_{j=1}^{p} \left(x_{i,j} - x_{i',j}\right)^2$$



Distance euclidienne entre paires de points :

	1	2	3	4	5	6
1	0.000	0.210	0.161	0.158	0.130	0.533
2	0.210	0.000	0.198	0.170	0.332	0.439
3	0.161	0.198	0.000	0.270	0.276	0.620
4	0.158	0.170	0.270	0.000	0.225	0.374
5	0.130	0.332	0.276	0.225	0.000	0.573
6	0.533	0.439	0.620	0.374	0.573	0.000

k-distance(A) : distance d'un objet A avec son k^{ème} voisin. Ex. 2-distance(6) = 0.439

$N_k(A)$: taille du voisinage d'ordre k de A. Attention, à cause des ex-aequo, $N_k(A) \geq k$

Distance atteignable (reachability distance) entre 2 objets : $rd_k(A,B) = \max\{k\text{-distance}(B), d(A,B)\}$

c.-à-d. Les points inclus dans le « rayon d'influence » d'ordre k de B sont considérés équivalents

Ex. 2-distance(4) = 0.170 → $rd_2(1,4) = \max\{0.170, 0.158\} = 0.170$

Ex. $rd_2(6,4) = \max\{0.170,0.374\} = 0.374$

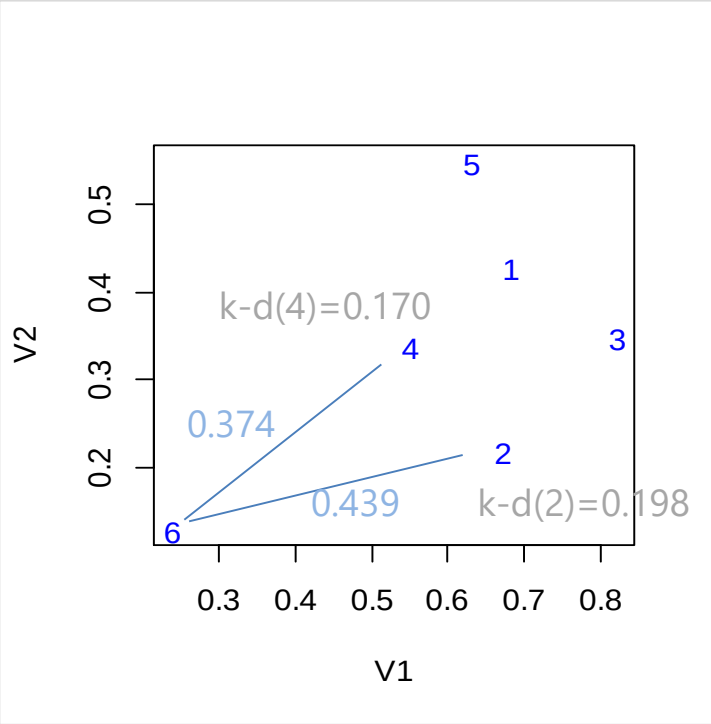
$rd_k(.\ ,\ .)$ n'est pas symétrique !

Quelques définitions et calculs (suite)

La densité locale (lrd : local reachability density) d'un point est l'inverse de la moyenne de son $rd_k(.,.)$ avec ses k-plus proches voisins

$$lrd_k(A) = \frac{1}{\frac{\sum_{B \in N_k(A)} rd_k(A,B)}{|N_k(A)|}}$$

$|N_k(A)|$ cardinal du k-voisinage de A (= k si pas d'ex-aequo)



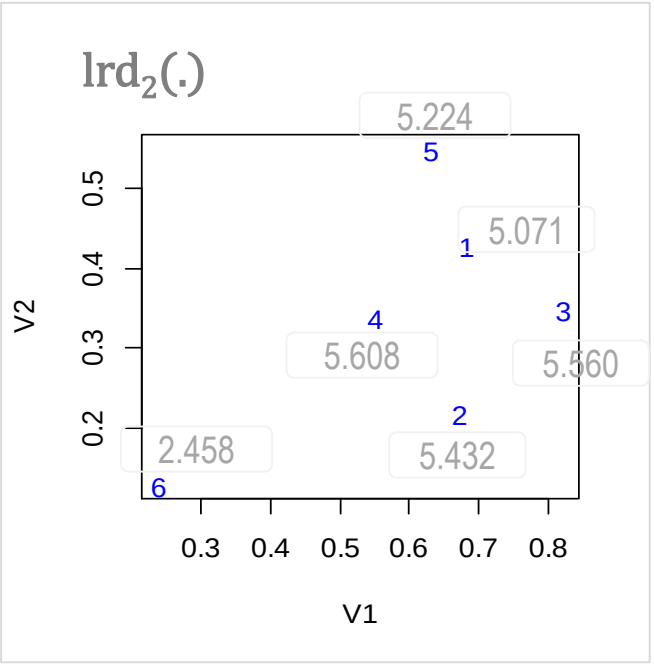
$rd_2(6,4) = \max\{0.170, 0.374\} = 0.374$

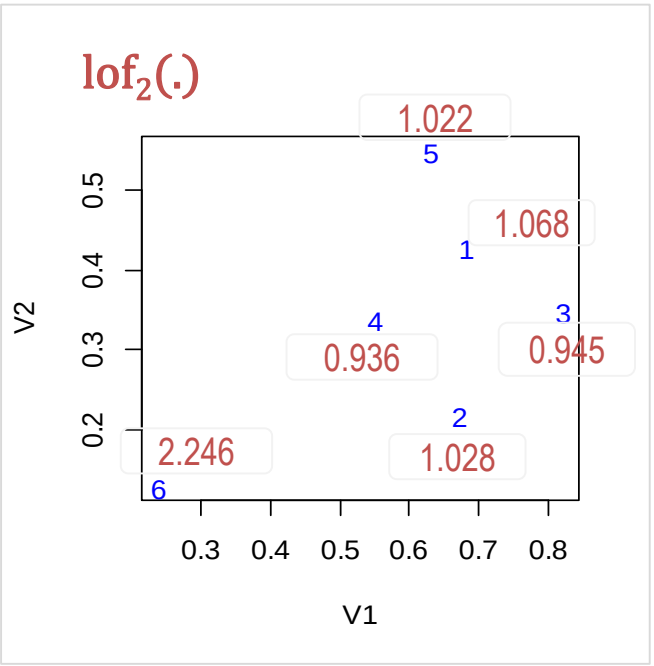
$rd_2(6,2) = \max\{0.198, 0.439\} = 0.439$

$$lrd_2(6) = \frac{1}{\frac{0.374 + 0.439}{2}} = 2.457756$$

$lrd_2(1) = 5.071$; $lrd_2(2) = 5.432$;
$lrd_2(3) = 5.560$; $lrd_2(4) = 5.608$;
$lrd_2(5) = 5.224$; $lrd_2(6) = 2.458$

On distingue les zones à forte densité !





Le facteur local d'anomalie (lof : local outlier factor) d'un point est obtenu en opposant sa densité locale avec celles de ses k-plus proches voisins

$$lof_k(A) = \frac{\sum_{B \in N_k(A)} \frac{lrd_k(B)}{lrd_k(A)}}{|N_k(A)|}$$

Exemple. $lof_2(6) = \frac{\frac{5.608}{2.458} + \frac{5.432}{2.458}}{2} = 2.246$

Règle de décision :

LOF ≈ 1, densité similaire à ses voisins

LOF < 1, densité plus élevée que ses voisins (*inlier*)

LOF > 1, densité moindre que ses voisins (*outlier*)

Le mieux toujours est de trier les données pour identifier les observations suspectes (LOF très élevé) !

LOF – Avantages et inconvénients

+

- Non dépendant au calcul toujours hasardeux d'un barycentre
- Approche locale, applicable même si les données sont organisées en clusters
- Tient compte de la densité des points
- Généralisable à d'autres mesures de distance (autre que euclidienne…)
- Peut travailler directement à partir d'une matrice de distance

-

- Interprétation du LOF difficile
- Valeur seuil de 1 discutable, mieux vaut identifier les décrochements
- Complexité des calculs, il faut une approche efficace de recherche des voisins

?

Comment fixer la valeur du paramètre k ?

- k faible, plus précis mais instabilité des résultats
- k fort, plus lissé mais risque de masquer les informations locales…

Identification des véhicules atypiques

UN EXEMPLE

Un ensemble de véhicules

Modele	Prix	Cylindree	Puissance	Poids	Conso
Daihatsu Cuore	11600	846	32	650	5.7
Suzuki Swift 1.0 GLS	12490	993	39	790	5.8
Fiat Panda Mambo L	10450	899	29	730	6.1
VW Polo 1.4 60	17140	1390	44	955	6.5
Opel Corsa 1.2i Eco	14825	1195	33	895	6.8
Subaru Vivio 4WD	13730	658	32	740	6.8
Toyota Corolla	19490	1331	55	1010	7.1
Ferrari 456 GT	285000	5474	325	1690	21.3
Mercedes S 600	183900	5987	300	2250	18.7
Maserati Ghibli GT	92500	2789	209	1485	14.5
Opel Astra 1.6i 16V	25000	1597	74	1080	7.4
Peugeot 306 XS 108	22350	1761	74	1100	9
Renault Safrane 2.2. V	36600	2165	101	1500	11.7
Seat Ibiza 2.0 GTI	22500	1983	85	1075	9.5
VW Golt 2.0 GTI	31580	1984	85	1155	9.5
Citroen ZX Volcane	28750	1998	89	1140	8.8
Fiat Tempra 1.6 Liberty	22600	1580	65	1080	9.3
Fort Escort 1.4i PT	20300	1390	54	1110	8.6
Honda Civic Joker 1.4	19900	1396	66	1140	7.7
Volvo 850 2.5	39800	2435	106	1370	10.8
Ford Fiesta 1.2 Zetec	19740	1242	55	940	6.6
Hyundai Sonata 3000	38990	2972	107	1400	11.7
Lancia K 3.0 LS	50800	2958	150	1550	11.9
Mazda Hachtback V	36200	2497	122	1330	10.8
Mitsubishi Galant	31990	1998	66	1300	7.6
Opel Omega 2.5i V6	47700	2496	125	1670	11.3
Peugeot 806 2.0	36950	1998	89	1560	10.8
Nissan Primera 2.0	26950	1997	92	1240	9.2
Seat Alhambra 2.0	36400	1984	85	1635	11.6
Toyota Previa salon	50900	2438	97	1800	12.8
Volvo 960 Kombi aut	49300	2473	125	1570	12.7

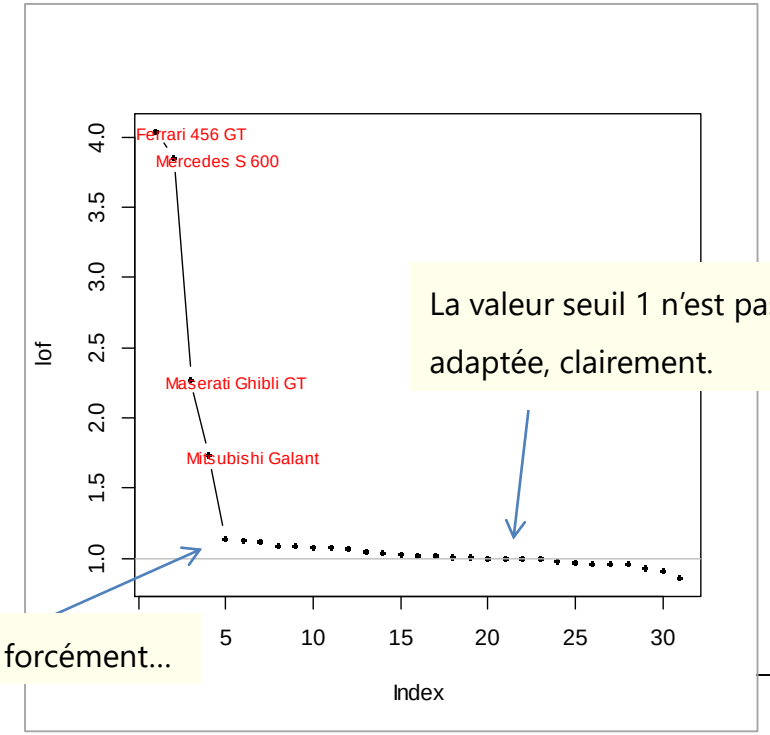
Comment identifier des véhicules atypiques – dont les caractéristiques se démarquent significativement des autres – dans cette base ?

Calculs sous R (1)

Les 3 premières on comprend, la Mitshubishi je doute, les autres (lof>1) pas bon.

```
#chargement du fichier
library(xlsx)
cars <- read.xlsx("cars_outliers.xlsx",header=TRUE,sheetIndex=1)
rownames(cars) <- cars$Modele
cars <- cars[-1]
print(str(cars))

#standardisation des données - important
Z1 <- scale(cars,center=TRUE,scale=TRUE)
print(Z1)

#identification des outliers
library(Rlof)
atyp <- Rlof::lof(Z1,k=3)
names(atyp) <- rownames(cars)
lof <- sort(atyp,decreasing=TRUE)
print(lof)

#décroissance du lof
plot(lof,type="b",pch=16,cex=0.5)
abline(a=1,b=0,col='gray')
text(3,lof[1],names(lof)[1],cex=0.75,col='red')
text(4.5,lof[2],names(lof)[2],cex=0.75,col='red')
text(5.5,lof[3],names(lof)[3],cex=0.75,col='red')
text(6.5,lof[4],names(lof)[4],cex=0.75,col='red')
```

```
Ferrari 456 GT       Mercedes S 600      Maserati Ghibli GT
      4.0367621           3.8476903               2.2698694
Mitsubishi Galant Fiat Tempra 1.6 Liberty   Opel Astra 1.6i 16V
      1.7300548           1.1390725               1.1264925
Fiat Panda Mambo L  Fort Escort 1.4i PT     VW Polo 1.4 60
      1.1188335           1.0924104               1.0848137
Honda Civic Joker 1.4 Toyota Previa salon   VW Golt 2.0 GTI
      1.0749252           1.0735969               1.0650363
Nissan Primera 2.0  Citroen ZX Volcane      Peugeot 306 XS 108
      1.0486296           1.0377882               1.0294415
Renault Safrane 2.2. V Ford Fiesta 1.2 Zetec Opel Omega 2.5i V6
      1.0182159           1.0176061               1.0126277
Lancia K 3.0 LS     Hyundai Sonata 3000     Mazda Hachtback V
      1.0051918           1.0022382               1.0022382
Volvo 960 Kombi aut Peugeot 806 2.0         Volvo 850 2.5
      1.0018507           0.9979525               0.9820399
Daihatsu Cuore      Suzuki Swift 1.0 GLS    Subaru Vivio 4WD
      0.9710589           0.9613885               0.9613885
Seat Alhambra 2.0   Opel Corsa 1.2i Eco     Toyota Corolla
      0.9582223           0.9300970               0.9102498
Seat Ibiza 2.0 GTI
      0.8616536
```



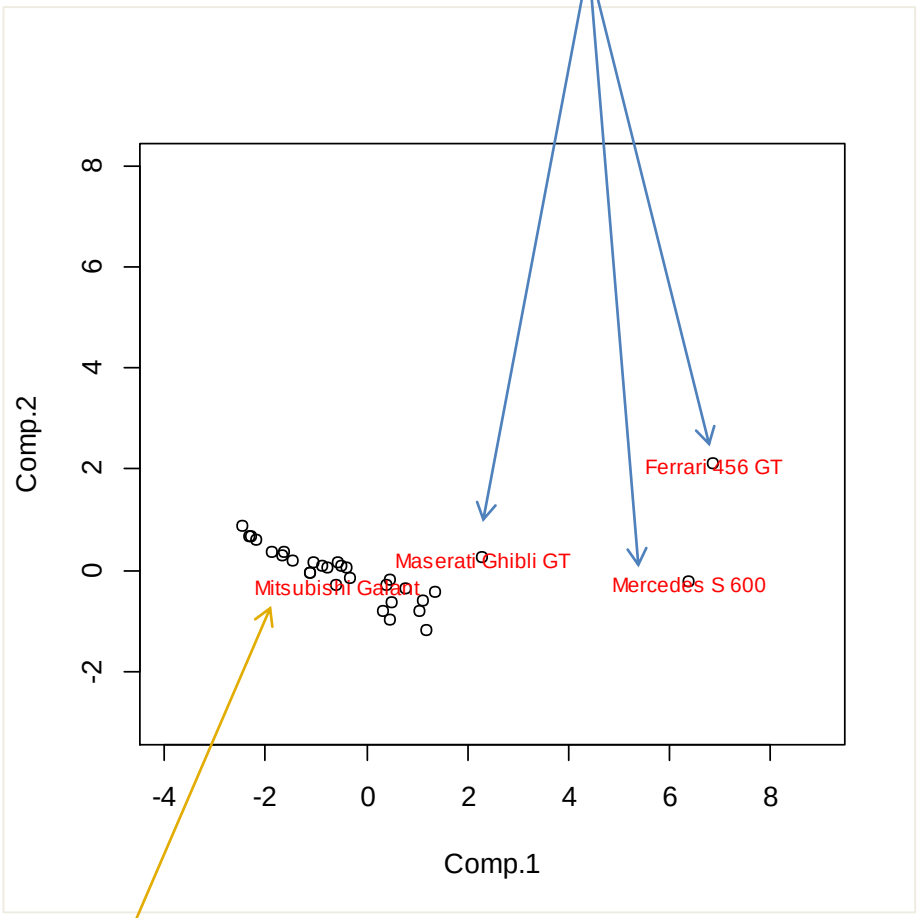
La valeur seuil 1 n'est pas adaptée, clairement.

Le « coude » interpelle, forcément…

Calculs sous R (2)

```
#analyse en composantes principales
acp <- princomp(cars,cor=TRUE,scores=TRUE)

#4 points atypiques : lof > 1.5
iza <- (atyp>1.5)
print(iza)

#projection dans le plan factoriel
plot(acp$scores[,1],acp$scores[,2],
     xlim=c(-3,8),ylim=c(-3,8), xlab='Comp.1',
     ylab='Comp.2',asp=1)
text(acp$scores[iza,1],acp$scores[iza,2],
     rownames(cars)[iza],col='red',cex=0.75)
```

L'analyse est confirmée



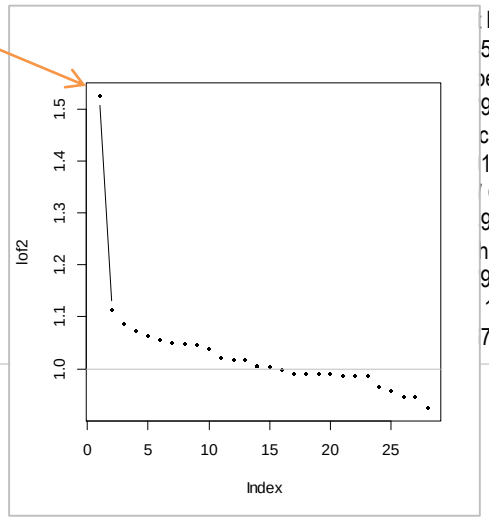
Reste un mystère à éclaircir (faux positif ?)

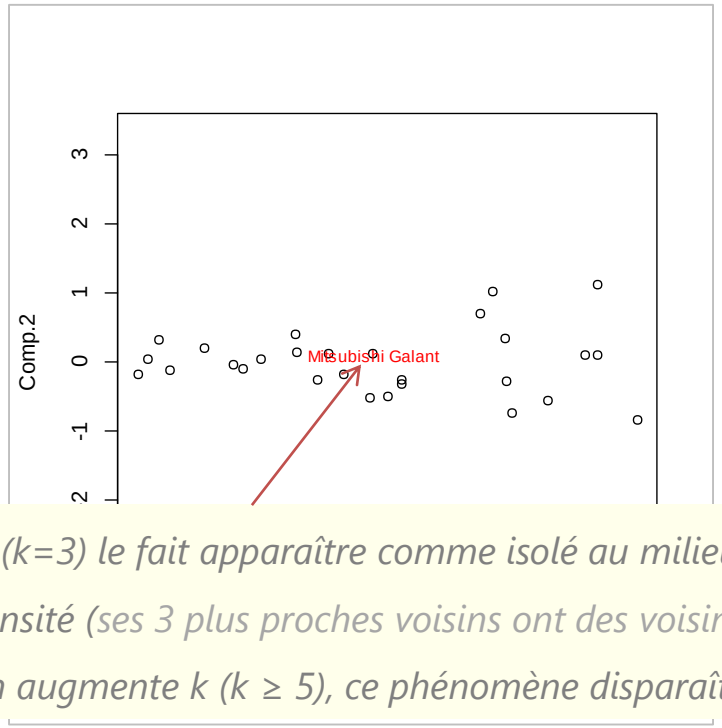
Calculs sous R (3)

Mitsubishi suspecte encore…

```
#refaire l'analyse sans les 3 atypiques
#(Mercedes, Ferrari, Maserati)
carsbis <- cars[atyp < 2.0,]

#centrage-réduction
Z2 <- scale(carsbis,center=TRUE,scale=TRUE)

#identification
atyp2 <- Rlof::lof(Z2,k=3)
names(atyp2) <- rownames(carsbis)
lof2 <- sort(atyp2,decreasing=TRUE)
print(lof2)

#graphique
plot(lof2,type="b",pch=16,cex=0.5)
abline(a=1,b=0,col='gray')
```

```
Mitsubishi Galant      VW Polo 1.4 60      Volvo 960 Kombi aut      Opel Omega 2.5i V6
        1.5256643           1.1131063                1.0861770               1.0723752
 Panda Mambo L    Opel Astra 1.6i 16V
570               1.0479916
en ZX Volcane    Hyundai Sonata 3000
959               1.0153769
cia K 3.0 LS     Toyota Previa salon
175               0.9971779
 Golt 2.0 GTI    Nissan Primera 2.0
906               0.9894906
nambra 2.0       Subaru Vivio 4WD
968               0.9658605
 1.2 Zetec       Toyota Corolla
732               0.9244434
```





Artefact : (k=3) le fait apparaître comme isolé au milieu d'une zone à forte densité (ses 3 plus proches voisins ont des voisins proches). Quand on augmente k (k ≥ 5), ce phénomène disparaît.

CONCLUSION

Local Outlier Factor - Conclusion

- La détection des anomalies a de nombreuses applications (identification des observations qui appartiennent à une autre population, des situations exceptionnelles, des comportements déviants [détection des intrusions par ex.], …)
- LOF est une approche non-supervisée locale basée sur le différentiel de densité entre les points d'un voisinage donné (nombre de voisins k est un paramètre)
- « Anomaly détection » peut–être « outlier détection » (sur la base étudiée) ou « novelty détection » (sur individus supplémentaires). LOF est applicable aussi en « novelty détection ».
- Le choix du paramètre (k) reste un problème ouvert…

RÉFÉRENCES

Références

- Wikipédia (en anglais) : « Outlier », « Anomaly detection », « Local outlier factor ».

- Documentation Scikit-learn 0.22, « Novelty and Outlier Detection », section 2.7.

- Breunig, Kriegel, Ng and Sander, « LOF: identifying density-based local outliers », Proc. of ACM SIGMOD – Int. Conf. on Management of Data, pp. 93-104, 2000.