Apprentissage Machine / Statistique

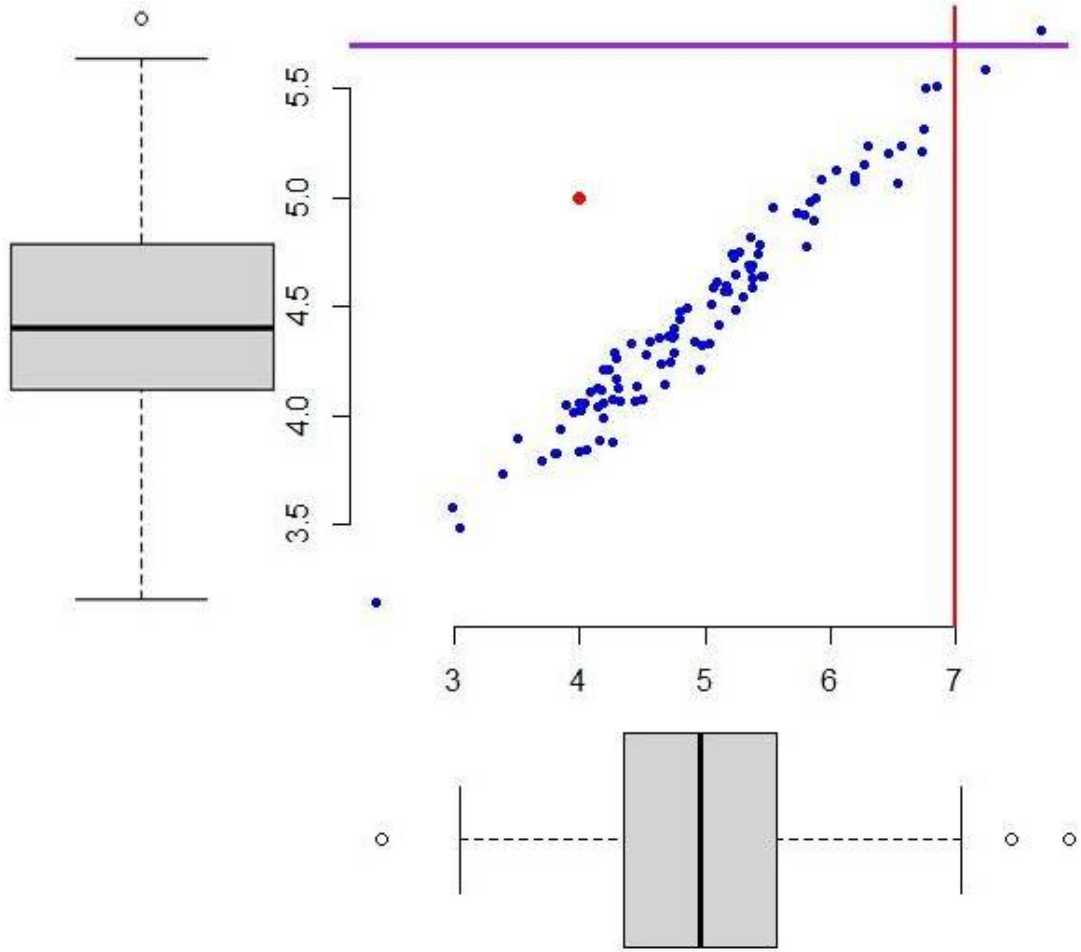
Détection d'anomalies

PHILIPPE BESSE

INSA – Dpt GMM
Institut de Mathématiques – ESP

Objectifs

- Historique : contrôle statistique des procédés
- Anomalie, défaut, intrusion, fraude, défaillance...
- Maintenance prédictive
- Définition uni *vs.* multi dimensionnelle
- Atypique (*outlier*) *vs.* “Normalité”
- Données atypiques *vs.* Valeurs extrêmes
- Approche synthétique d’une vaste bibliographie



Point atypique : uni, bi-dimensionnel ou multi...

Taxonomie schématique

- Choix partiel et partial. Voir la bibliographie pour :
  - Les approches unidimensionnelles (SPC)
  - Celles multidimensionnelles paramétriques (gaussiennes)
- Méthodes non paramétriques multidimensionnelles
  - Historique d'anomalies
  - Méthodes Supervisées : cf. épisodes précédents
    Attention classes déséquilibrées
  - Méthodes Non supervisées :
    outliers, one class classification (OCC), novelty detection

Méthodes non supervisées

- Principe : Même dans le cas non-paramétrique, une *anomalie est définie par rapport à un modèle*
  - paramétrique généralement gaussien
  - Non paramétrique dépendant des données :*e.g.* voisinage local ou k plus proches voisins
  - Explicite si variable cible Y : régression ou classification
  - Implicite : densité de probabilité
- Méthodes pour variables quantitatives ou mixtes

Cas non supervisé paramétrique

- Relatif à une variable cible : distance de Cook
- Sans variable cible
  - Test de Hotteling (1931)
  - Densité multidimensionnelle et Distance (D_M) de Mahalanobis $(\mathbf{S}^{-1})$
  - ACP (Ruiz Gazen et Caussinus 2007)
    - Estimation robuste de la matrice de covariance de $\mathbf{X}_{(n,p)}$

$$
\begin{aligned}
\mathbf{S} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})' \\
w_i &= \exp(-\beta/2||\mathbf{x}_i-\bar{\mathbf{x}}||^2_{\mathbf{S}^{-1}}) \\
\mathbf{R} &= \frac{\sum_{i=1}^{n} w_i(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})'}{\sum_{i=1}^{n} w_i}
\end{aligned}
$$

    - ACP avec métrique $\mathbf{M}=\mathbf{R}^{-1}$

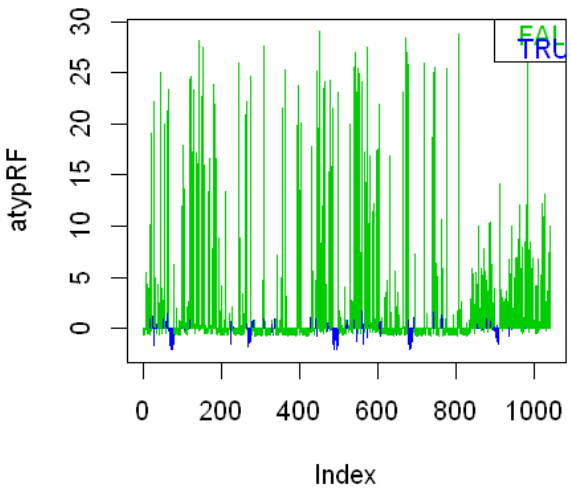
Cas non supervisé non paramétrique

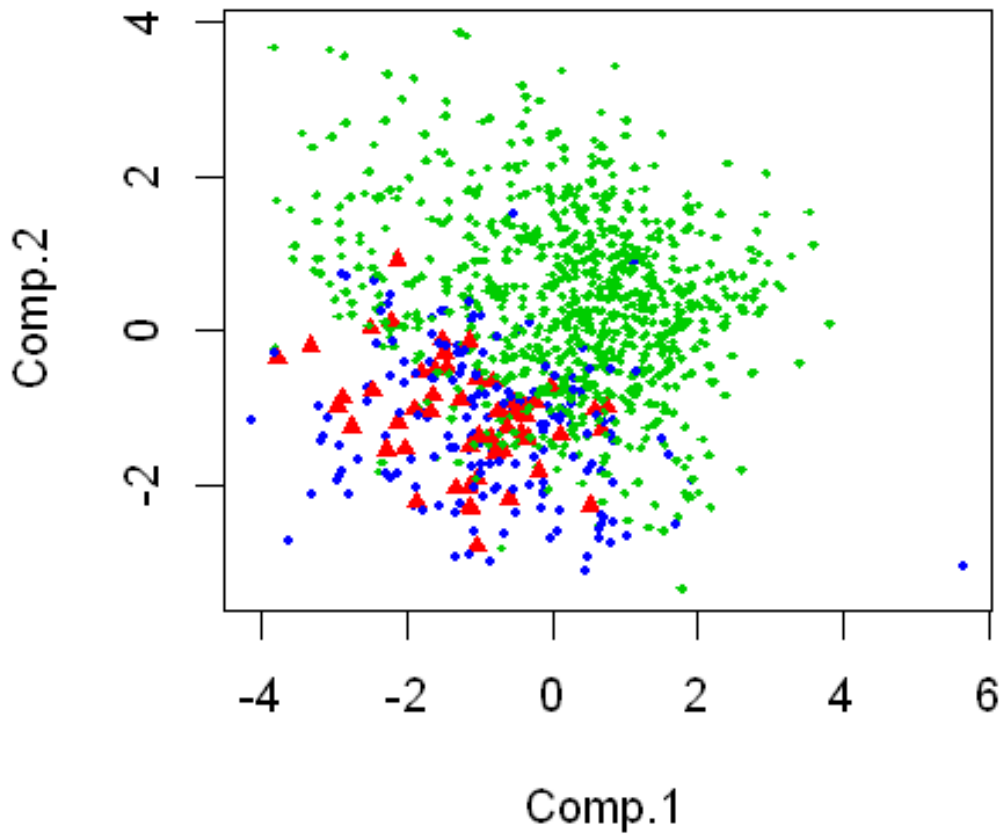
- Relatif à une variable cible : *random forest* (Breiman 2001)
- Sans variable cible : vaste littérature (Aggarwal 2017)
  - Méthodes de classification non supervisée : k-means, DBSCAN, CAH...
  - LOF et avatars (GLOSH...), OCC SVM, OCC RF, *isolation forest*...

Anomalies par rapport à une forêt aléatoire

- $Y = f(\mathbf{X})$ estimé par *random forest*
- Similarité entre deux observations :
  nb moyen d'occurrences dans la même feuille
- Score d'anomalie

- P : somme des carrés des proximités des observations à celles de sa classe
- Score : Effectif de la classe divisé par P puis "centré" par la médiane et "réduit" par MAD (médiane des écarts absolus à la médiane).





Ozone : atypiques au sens de RF expliquant la concentration.

Local Outlier Factor (LOF)

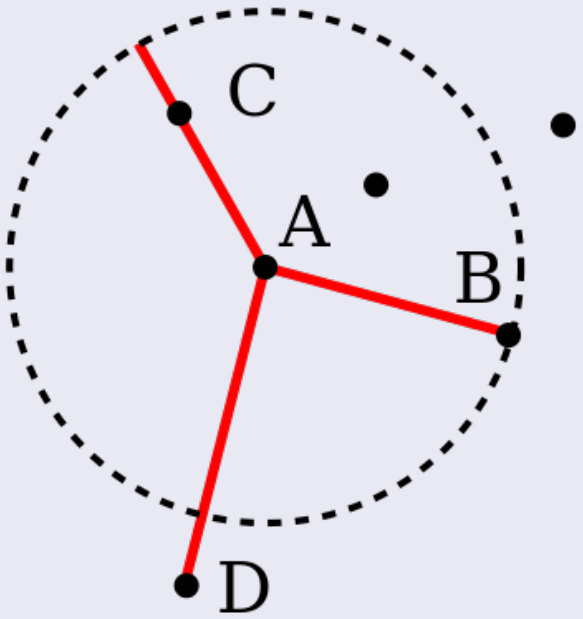
- Densité locale et proximité des points du k voisinage
- Choix d'une distance : l_1, l_2, matrice $n \times n$...
- *LOF* découle de DBSCAN
- $\text{Dist}_k \approx$ distance du k-ième plus proche voisin

$$V_k(x) = \{x \in D \quad | \quad d(x,y) \leqslant \text{Dist}_k(x)\}$$

$$\text{RDist}_k(x,y) = \max\{\text{Dist}_k(y), d(x,y)\}$$

$$\text{LRDens}(x) = 1/\left(\frac{\sum_{y \in V_k(x)} \text{RDist}_k(x,y)}{\text{card}(V_k(x))}\right)$$

$$\text{LOF}_k(x) = \frac{\sum_{y \in V_k(x)} \frac{\text{LRDens}(y)}{\text{LRDens}(x)}}{\text{card}(V_k(x))}$$



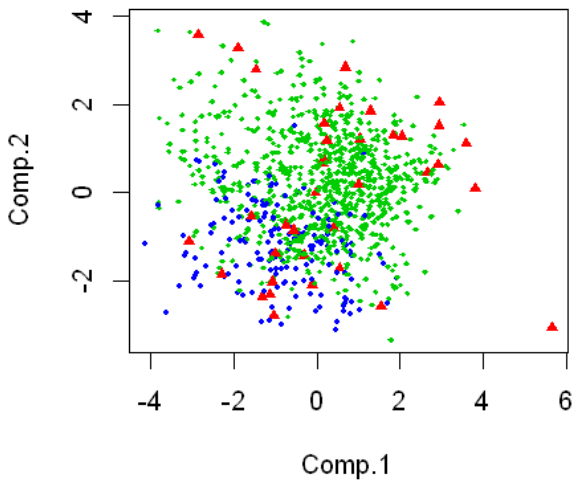
- *LoOF, IUOS, GLOSH...*

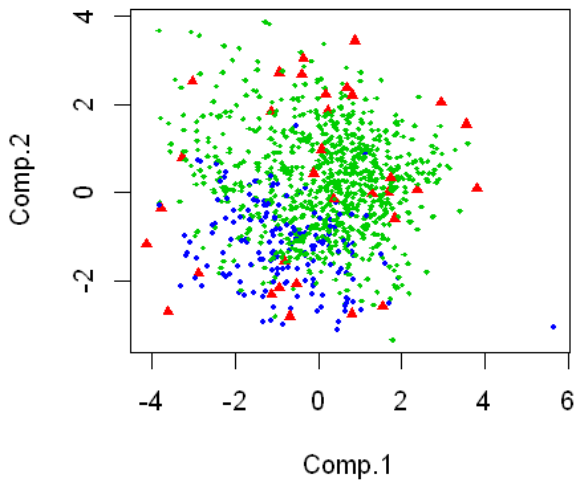
One Class SVM (OCC SVM)

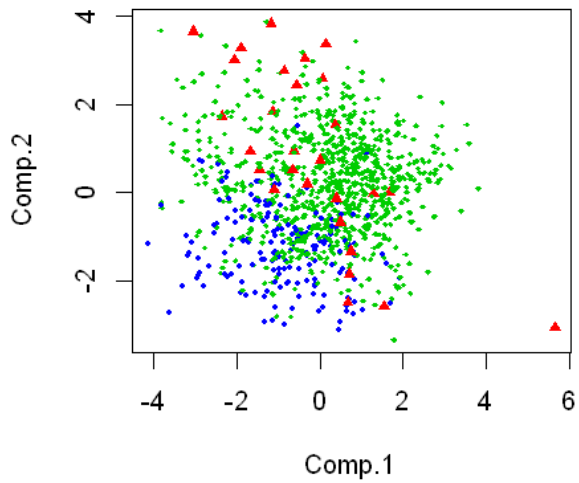
- Enveloppe ou support des observations suffisamment denses
- Séparer les observations de l'origine
- Hyperplan dans l'espace intermédiaire (*feature space*
- Capter la zone dense des données

One Class random forest (OCC RF)

- Forêt aléatoire non supervisée de `randomForest` :
  - Générer deux classes d'observations
  - Observées et permutations aléatoires des variables
  - Forêt qui discrimine ces deux classes
  - Matrice des *proximités* des points 2 à 2
  - Matrice de distances puis CAH ou MDS
- Score d'anomalie issu des proximités







Ozone : atypiques selon LOF, OCC SVM, OCC RF

Isolation forest (`Scikit-learn`)

- *Isolation tree*
- B (100) arbres sur sous-échantillon (256)
- Score : Moyenne des longueurs des chemins

Conclusion ?

- Problème complexe et bibliographie explosive
- Agréger, conjuguer plusieurs critères
- Comment choisir ?
  - Méthode, seuil ou sensibilité à régler
  - Même si non supervisé : nécessité d'un historique
- Données complexes : signaux, courbes, images, chemin...
- Thèses en cours ou à venir