Introduction aux Big Data

Julien JACQUES

Rétrospective historique (1/4)

- Historiquement, les premières approches statistiques étudient un **petit nombre n d'individus décrits par un petit nombre p de variables**. Ces données sont issues de plans d'expériences.

Rétrospective historique (2/4)

- **1990s (MO)** : les entreprises commencent à stocker de plus en plus de données concernants leur clients, sans planification expérimentale. Les méthodes statistiques classiques sont massivement utilisées pour extraire de la connaissance de ces données (CRM, gestion de la relation client). C'est la naissance du **data mining**.

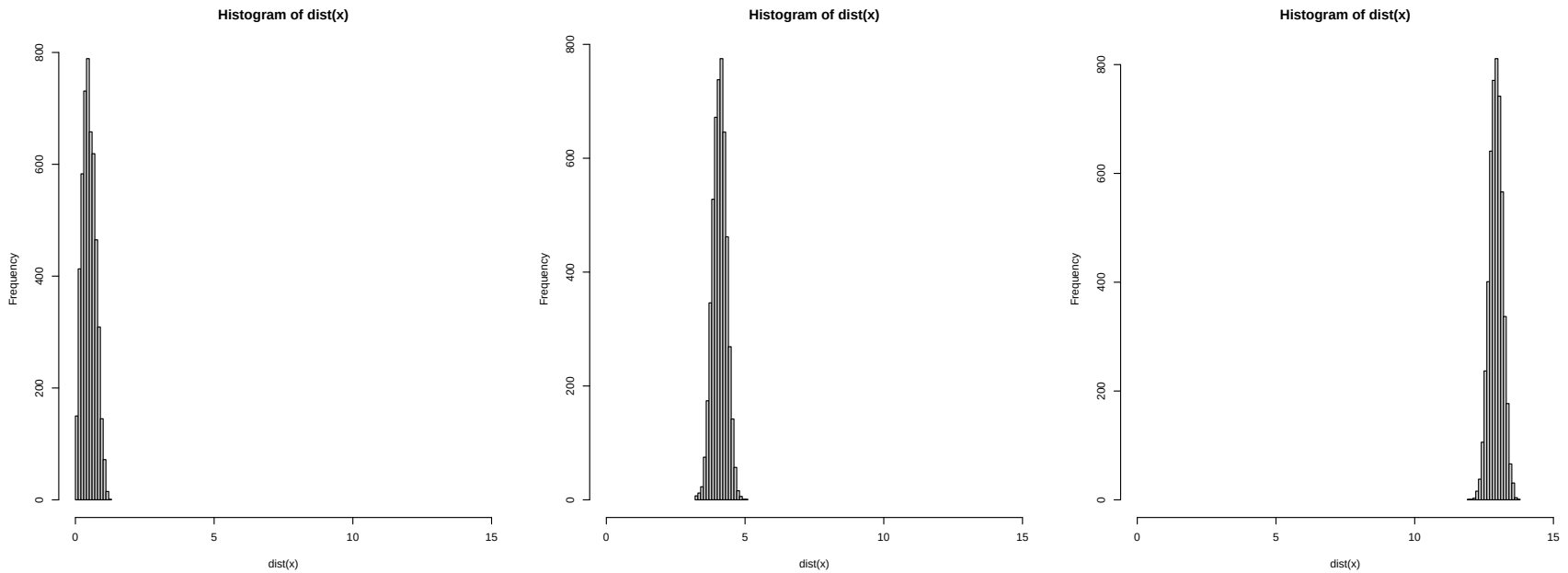
Rétrospective historique (3/4)

- **2000s (GO)** : **première révolution** du data mining avec l'avénement de la bioinformatique et des données omiques : on observe beaucoup de variables sur peu d'individus ($n << p$). On parle du **fléau de la dimension** et on doit réduire la dimension et développer de nouvelles **méthodes parcimonieuses**.

Fléau de la dimension

Illustration des distances entre points choisis uniformément sur $[0,1]^p$

```
par(mfrow=c(1,3))
for (p in c(2,100,1000)){
  x=matrix(runif(100*p),ncol = p)
  hist(dist(x),xlim=c(0,15))
}
```



En grande dimension, **l'espace est vide**, tous les points sont loins les uns des autres, aucune observation ne ressemble à aucune autre

Rétrospective historique (4/4)

- **2010s (TO)** : **seconde révolution** due au développement d'internet (commerce en ligne, réseau sociaux). On parle de **big data** et de science des données.
  Les big data sont des données :
  - **volumineuses** : à la fois en nombre n d'observations et en nombre p de variables
  - de nature **variée** : quantitative, qualitative, texte, image, réseau...
  - **évolutive** : les jeux de données grandissent en continue. On parle de flux de données

Quelques lectures intéressantes

Philippe Besse, Aurélien Garivier, Jean-Michel Loubes. Big Data Analytics - Retour vers le Futur 3; De Statisticien à Data Scientist. https://hal.archives-ouvertes.fr/hal-00959267

Philippe Besse, Nathalie Vialaneix. Statistique et Big Data Analytics; Volumétrie, L'Attaque des Clones. 2014. https://hal.archives-ouvertes.fr/hal-00995801v3

Big Data : des données volumineuses

Un grand nombre n d'observations :

- théoriquement, plus n est grand, mieux les modèles seront estimés. C'est donc *statistiquement* un avantage...
- ... sauf si on ne peut plus mettre en mémoire le jeu de données
- il faut alors faire appel à des architectures big data qui distribue les calculs sur différentes machine (Hadoop, Spark)
- l'enjeu est alors de savoir estimer un modèle en **distribuant les données** sur différentes machine (Map Reduce)

Nous n'aborderons pas ceci dans cette formation

Big Data : des données volumineuses

Un grand nombre p de variables :

- là par contre, c'est théoriquement plus complexe...
- certains modèles ne peuvent être estimés

```
x=matrix(rnorm(80),8,10)
y=x %*% 1:10
lm(y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)           x1           x2           x3
##      1.4880       5.8685       4.7659      -0.6869
##          x6           x7           x8           x9
##     24.0514       9.9607           NA           NA
```

Big Data : des données volumineuses

Un grand nombre p de variables :

- plus il y a de variables, plus il y a de chances de voir des variables corrélées
  - ces variables peuvent être réellement corrélées
  - ou une corrélation fictive peut apparaitre lorsque $n << p$

```
cor(matrix(rnorm(18),3,6))
```

```
##              [,1]       [,2]       [,3]        [,4]
## [1,]  1.0000000  0.7119641 -0.5812057 -0.13011509  0.930
## [2,]  0.7119641  1.0000000  0.1576353  0.60360897  0.919
## [3,] -0.5812057  0.1576353  1.0000000  0.88246251 -0.243
## [4,] -0.1301151  0.6036090  0.8824625  1.00000000  0.240
## [5,]  0.9309473  0.9192177 -0.2439257  0.24091899  1.000
## [6,] -0.9895580 -0.8057440  0.4578455 -0.01415372 -0.973
```

On parle de **statistique en grande dimension**

Nous aborderons ceci dans cette formation, en commençant par le cas de la régression

Big Data : des données (variables) de nature variée

Cas des variables quantitatives et catégorielles :

- bien avant les big data, on a eu à faire à des données de ce type
- deux approches sont alors possibles :
  - utiliser des méthodes qui nativement peuvent travailler avec des variables mixtes (Trees, RandomForest)
  - uniformiser la nature des données, en transformant les variables catégorielles en variables quantitatives (l'inverse est à proscrire, cela ferait perdre trop d'information) :
    - en variable binaire indicatrices de catégories (attention, la dimension explose. . .)
    - via une ACM et en travaillant sur les composantes principales (on perd en interprétation)

Nous aborderons ceci dans cette formation, en commençant par le cas de la classification

Big Data : des données (variables) de nature variée

Cas des données textuelles :

- représentation **bag-of-words** : à chaque mot du dictionnaire on associe une variable discrète qui compte le nombre d'occurences du mot dans le texte. Attention : une telle représentation est en très grande dimension et contient essentiellement des 0
- des techniques modernes (**Word2Vec, Bert**) plongent les mots dans des espaces vectoriels de dimension réduite, de sorte que deux mots proches ayant un sens proche soient proches dans cet espace

Cas des images :

- une image n'est qu'un ensemble de pixel, un pixel étant décrit par une ou des variables quantitatives :
  - niveau de gris $\in [0, 255]$ pour les images en noir et blanc
  - niveau de rouge, vert et bleu $\in [0, 255]^3$ pour les images en couleur

Nous travaillerons avec des images N&B dans cette formation

Big Data : des données (variables) de nature variée

Travailler avec tous ces types de données ensemble pose encore des soucis, même si on sait à chaque fois se ramener à des représentations quantitatives :

- la dimension induite devient très grand
- le poids de chaque type de variables est difficile à gérer

Nous sommes sur des problématiques de recherche actuelle en statistique et machine learning. . .

Big Data : des données évolutives

- Dans certains applications, les données arrivent par flux régulier.
- Recommencer l'étude à chaque fois est coûteuse en temps notamment
- Certains algorithmes *online* permettent de mettre à jour les paramètres des modèles au fur et à mesure

Ce point ne sera pas abordé dans cette formation, et là encore nous sommes sur des problématiques de recherche actuelle