Statistiques et données massives

Joseph Saint Pierre

Rencontre ingénieurs statisticiens 27 janvier 2014

Les origines de ce petit exposé sont nombreuses et pour certaines anciennes ou très récentes.

Le 20 juillet 2013 j'ai écouté l'émission « place de la toile »sur France Culture, voici le lien de l'émission :

`http://www.franceculture.fr/emission-place-de-la-toile-entretien-avec-chris-anders`

Bonjour.

J'ai beaucoup apprécié l'émission et je me suis senti très concerné par le dernier quart d'heure de l'entretien avec Chris Anderson où il a été question de l'article de 2008 sur les données massives : `http://www.wired.com/science/discoveries/magazine/16-07/pb_theory`. Un léger reproche à la traduction certaines ne sont pas traduites comme les logiciels "sources ouvertes" ou "données massives" qui restent "open source" et "big data".

En tant que mathématicien, statisticien je suis en désaccord avec le point de vue implicitement défendu par l'invité dans son article cité plus haut. Je suis beaucoup plus proche de la position défendue ici : `http://www.guardian.co.uk/news/datablog/2012/mar/09/big-data-theory` Il me semble simpliste d'opposer une démarche statistique fondée sur les corrélations à une approche de modélisation plus évoluée. Si les capacités à stocker des données a augmenté rapidement les capacités de calcul ont simultanément augmenté et il est possible d'améliorer les modèles. Au fur et à mesure que les volumes de données sont devenus importants les méthodes de calcul ont évolué et elles continuent d'évoluer.

Je reconnais volontiers que je défends un point de vue corporatiste de mathématicien défendant une position pour défendre mon emploi...

Très cordialement

Quelques petits points sur ma page web.

J'ai commencé à avoir une page web personnelle en 1994, il n'y avait qu'une photographie et c'était à titre de test, je considérais cela comme très éloigné de mon centre

d'intérêt c'est à dire des statistiques.

C'est en 1999 que ma page web à commencé à se remplir avec des documents anciens grâce à la commande latex2html.

C'est vers 2003 que je me suis rendu compte qu'en regardant des fichiers logs on pouvait voir des liens, des accès par des moteurs de recherche. J'ai commencé à archiver quelques accès remarquables, notamment la documentation d'initiation à R écrite en 2002.

À partir de 2006 j'ai commencé à archiver plus de choses et surtout à voir le succès de la documentation de PSPP.

C'est uniquement en 2013 que l'idée m'est venue de voir les « corrélations »entre les téléchargements.

Depuis novembre je teste une nouvelle version de R sur une nouvelle machine. Sur cette nouvelle version de R j'ai vu qu'il était possible de produire des graphes et je sais que cela peut intéresser des statisticiens et les données sur les téléchargements étaient disponibles pour produire un graphe auquel je peux donner quelques interprétations.

Mon idée d'exposé est de montrer le lien entre trois expsés que j'ai déjà présentés dans le cadre des rencontres d'ingénieurs statisticiens.

15 Octobre 2012
Le système unix et statistiques

8 mars 2012
PSPP une alternative à SPSS ?

11 Juillet 2013
Les statistiques chez Wikipedia

Les commandes unix grep, uniq, cut etc. servent à garder les lignes et les champs pertinents.

PSPP sert à mettre les données en forme et faire des petits calculs ;

Les liaisons entre les pages sont calculées à travers le test exact de Fisher (page wikipedia). PSPP et R.

Ce test détermine les arêtes du graphe si test significatif arête, sinon pas d'arête.

Le but de l'exposé n'est pas de faire de la recherche il s'agit essentiellement de voir les utilisations enchainées de logiciels et du système unix avec peut être des macros emacs... La démarche est importante mais la mise en oeuvre essentielle. C'est un tout petit exposé pour des ingénieurs statisticiens.

Voici une ligne extraite d'un fichier d'accès à ma page web. On peut voir le numéro

IP de l'ordinateur (plaquette, téléphone etc...) qui a accédé, la date, l'heure, le document accédé, le logiciel utilisé, le système etc.

```
81.252.107.233 - - [08/Jan/2014:12:52:36 +0100] "GET
/~stpierre/doc-pspp.pdf HTTP/1.1" 206 65536
"http://lists.gnu.org/archive/html/pspp-dev/2006-07/msg00081.html"
"Mozilla/5.0 (Windows NT 5.1; rv:26.0) Gecko/20100101 Firefox/26.0" "-"
```

On obtient ce genre de ligne avec une commande unix grep du genre :

```
grep stpierre access_log | grep pdf | grep -iv bot
```

Avec la commande unix cut on peut sélectionner des informations
La commande :

```
cut -f1,4,5,7,11 -d" "
```

Fournit le résultat suivant :

```
81.252.107.233 [08/Jan/2014:12:52:36 +0100] /~stpierre/doc-pspp.pdf
"http://lists.gnu.org/archive/html/pspp-dev/2006-07/msg00081.html"
```

La commande :

```
cut -f1,4,5,7 -d" "
```

donne :

```
81.252.107.233 [08/Jan/2014:12:52:36 +0100] /~stpierre/doc-pspp.pdf
```

Puis avec un petit nettoyage, toujours avec des commandes unix, à une information plus réduite :

```
81.252.107.233 [08/Jan/2014:12:52:36] doc-pspp
```

voici un extrait du fichier constitué par ces commandes.

```
122.178.192.17 [29/Dec/2013:17:40:40] doc-pspp
185.29.166.230 [29/Dec/2013:20:42:33] doc-pspp
89.70.62.170 [29/Dec/2013:21:40:57] doc-pspp
112.123.168.54 [29/Dec/2013:22:56:19] doc-pspp
86.73.42.47 [30/Dec/2013:05:48:09] wiki-stats-notes
82.228.62.111 [30/Dec/2013:09:36:42] glim-genmod
212.1.215.175 [30/Dec/2013:09:39:13] doc-pspp
195.101.137.28 [30/Dec/2013:09:45:02] glim-genmod
85.69.156.192 [30/Dec/2013:10:14:44] thesen
```

À partir de fichier il est possible de faire des statistiques simples sur le nommbe de fois où un document est accédé, les nombres d'accès par jour, semaine, mois, année, par numéro IP etc. Avec des tableaux croisés :

	doc-R	doc-pspp
2007	1948	896
2013	1076	2232

Il est possible de se demander si il existe une liaison entre les documents accédés et cela demande une petite manipulation statistique. Pour cela j'ai utilisé le logiciel PSPP.

```
cross ip by doc
 /cell=count
 /stat=none
```

Compte tenu de la taille du fichier il a été préférable d'utiliser des commandes AGGREGATE.

Voici le programme complet de pspp permettant de transformer le fichier texte en tableau de contingence.

```
data list list file=ddin
 /ip (A18) doc (A50) .
autorecode var = doc into docn  .
select if (docn =10)
 or (docn=11) or (docn=14)  or (docn=15)  or (docn=18)  or (docn=29)
 or (docn=32) or (docn=33)  or (docn=35)  or (docn=37)  or (docn=38)
 or (docn=40) or (docn=41)  or (docn=46)  or (docn=48)  or (docn=50)
 or (docn=51) or (docn=52).
autorecode var= docn into docn2.
autorecode var=ip into ipn .
if (docn2=1) doc1=1.
if (docn2=2) doc2=1.
if (docn2=3) doc3=1.
if (docn2=4) doc4=1.
if (docn2=5) doc5=1.
if (docn2=6) doc6=1.
if (docn2=7) doc7=1.
if (docn2=8) doc8=1.
if (docn2=9) doc9=1.
if (docn2=10) doc10=1.
if (docn2=11) doc11=1.
if (docn2=12) doc12=1.
if (docn2=13) doc13=1.
if (docn2=14) doc14=1.
if (docn2=15) doc15=1.
if (docn2=16) doc16=1.
if (docn2=17) doc17=1.
if (docn2=18) doc18=1.
AGGREGATE OUTFILE=* /BREAK=ipn
 /page1=sum(doc1)  /page2=sum(doc2)  /page3=sum(doc3)  /page4=sum(doc4)
 /page5=sum(doc5)  /page6=sum(doc6)  /page7=sum(doc7)  /page8=sum(doc8)
 /page9=sum(doc9)  /page10=sum(doc10)  /page11=sum(doc11)  /page12=sum(doc12)
 /page13=sum(doc13)  /page14=sum(doc14)  /page15=sum(doc15)  /page16=sum(doc16)
 /page17=sum(doc17) /page18=sum(doc18).
recode page1 to page18 (0=0) (else=1).
formats page1 to page18 (F2.0) .
write outfile=ddout
 /page1 to page18 .
execute .
```

À partir du fichier précédent ces commandes permettent de créer une table de contingence, avec en ligne les numéros IP 34973 et en colonne les documents 18. il est inutile de garder le numéro IP pour les traitements suivants. Il est aussi inutile de garder les doublons et on obtient donc un tableau de 0 et de 1 dont voici un extrait :

0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

Avec ce tableau par des produits de colonnes il est très facile de constituer les tables de contingence deux à deux, il y en $17 * 18/2 = 153$.

Le produit simple de deux variables, suivi d'un calcul de la somme de la variable obtenue fournit le nombre de numéros IP qui ont accédés aux deux documents spécifiés.

```
compute v12_18=v12*v18 .
descriptive var =v12_18
 /stat=sum.
```

Il devient, avec ce nombre facile, de reconstituer le reste des tables de contingence qui n'ont qu'un seul degré de liberté.

Avec les valeurs de la table de contingence on peut obtenir avec R le calcul du test exact de Fischer :

$fisher.test(matrix(c(4, 168, 176, 34625), nrow = 2))$

Voilà ce que fournit R pour la première table :

```
        Fisher's Exact Test for Count Data

data:  matrix(c(4, 168, 176, 34625), nrow = 2)
p-value = 0.01224
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
```

```
  1.247741 12.428184
sample estimates:
odds ratio
  4.683567
```

Il y a 153 tests. Beaucoup trop de tests sont significatifs avec un seuil usuel de 5% en retenant un seuil beaucoup plus restrictif on obtient 53 arêtes

Pour construire des graphes, j'ai utilisé R et la librairie igraph.

```
library(igraph)
madj<-matrix(c( 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1,
1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0,
1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1,
0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1,
0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1,
0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0,
1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1,
1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0,
0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0,
1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0,
1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 1, 0), nrow=18, ncol=18, byrow=TRUE, dimnames =
list(c("aggl", "arbr", "dich", "docR", "pspp", "expl", "glim", "hist",
"limt", "llcc", "mash", "mmh", "mdl","nrn", "cls", "smps", "tho",
"thn"),c("aggl", "arbr", "dich", "docR", "pspp", "expl", "glim",
"hist", "limt", "llcc", "mash", "mmh", "mdl","nrn", "cls", "smps",
"tho", "thn")))

g<-graph.adjacency(madj, mode="undirected")

plot (g)
```

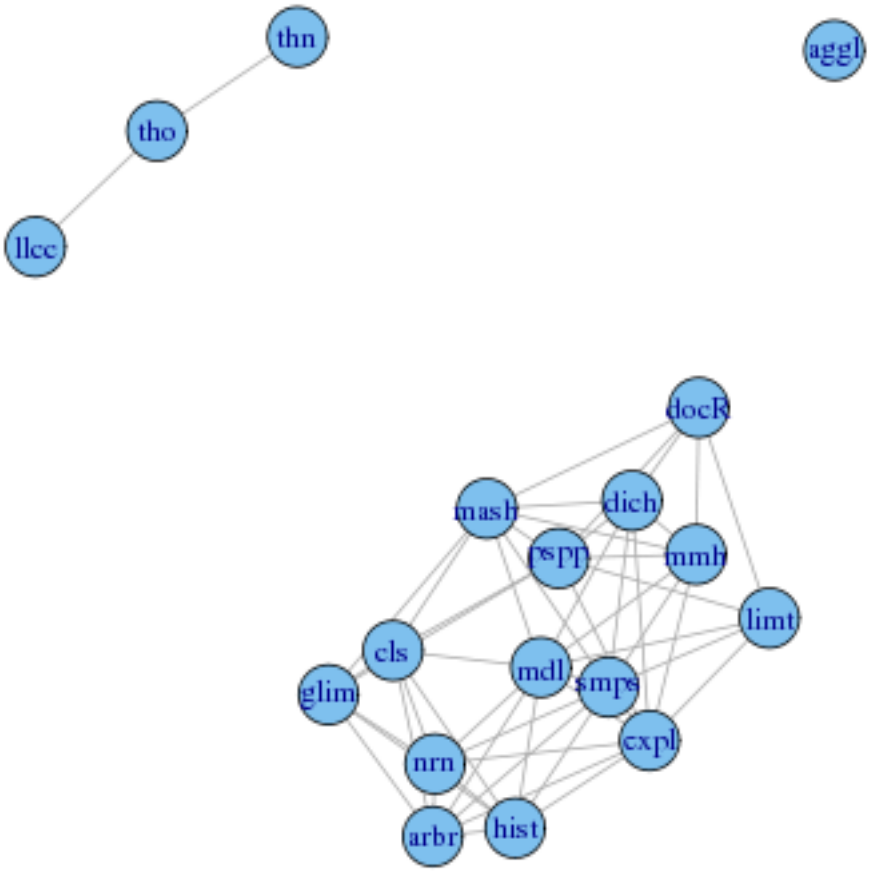

FIG. 1 – Graphe 2

Cela produit un graphe qu'il peut être intéressant d'analyser. igraph dans R permet de produire énormément de critères pour décrire les graphes, on peut notamment obtenir le PageRank http ://fr.wikipedia.org/wiki/PageRank

L'analyse des liaisons permet d'imaginer un système de proposition analogue à ceux que l'on trouve sur Internet, pour suggérer des documents aux internautes.

Avec un niveau de test différent on a plus d'arêtes dans le graphe et voici le dessin du graphe.

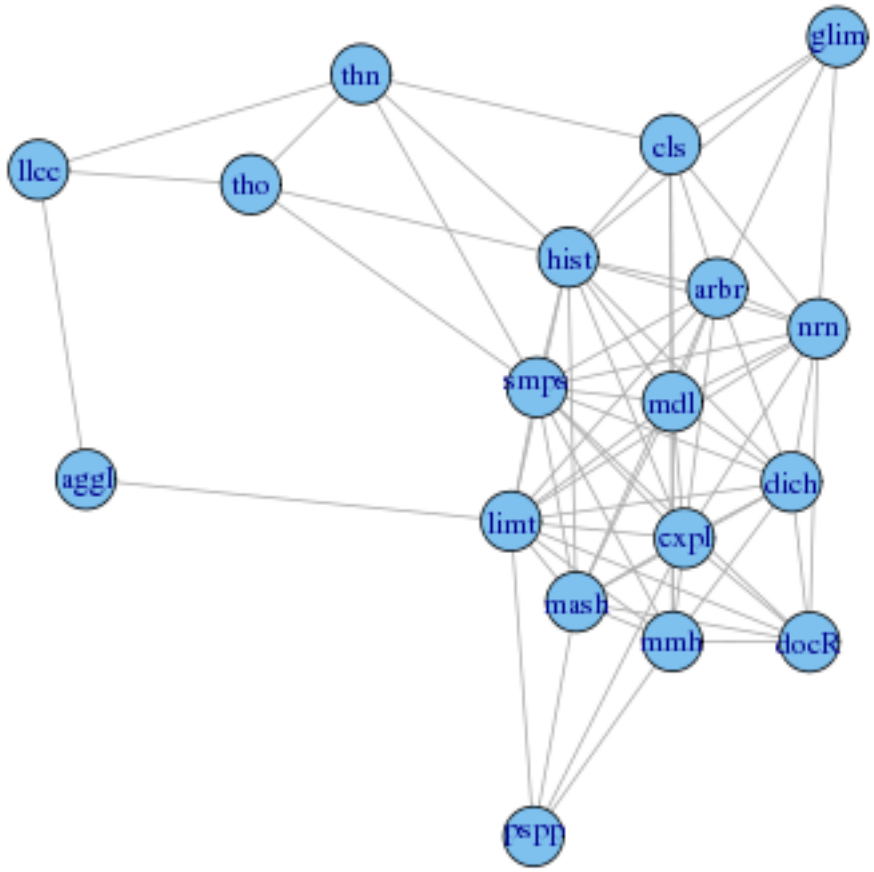

FIG. 2 – graphe 2

Voici un extrait d'un exposé que j'avais écrit en 1999 sur le croisement systématique de variables binaires pour détecter des liaisons.

Il existe une méthode assez simple pour détecter les liaisons entre modalités de variables qualitatives qui consiste à croiser systématiquement les variables entre elles. Il est fastidieux et long d'analyser d'énormes quantités de tableaux croisés Dans le logiciel SPAD il existe une commande DEMOD (pour description de modalités) qui permet de rationaliser ce travail en ne gardant que les liaisons qui sont significatives au sens d'un certain test. Dans les anciennes versions de SPAD, celle que j'ai commencé à utiliser en 1984 par exemple, la procédure s'appelait TAMIS, un certain nombre de statisticiens

appellent encore ce principe de croisement par le nom ancien de cette procédure. Ce principe de croisement se ramène toujours à des croisements de variables à deux modalités. le logiciel commencerait par créer virtuellement autant de variables que de modalités en jeu dans la procédure de description, chacune de ces variables est binaire présence contre absence. SPAD utilise la loi hyper-géométrique qui est plus justifiée dans le cas des petits effectifs. La loi hyper-géométrique est celle sur laquelle se fonde le test exact de Fischer.

Un autre extrait d'un autre exposé ancien :

Statistique n'est pas probabilité. Sous le nom de statistique mathématique des auteurs, qui, je vous le dis en français n'écrivent guère dans notre langue, ont édifié une pompeuse discipline riche en hypothèses qui ne sont jamais satisfaites. (Jean-Paul Benzécri, l'Analyse Des Données, 1972). Benzécri est l'inventeur de la méthode qui s'appelle l'analyse factorielle des correspondances (AFC) et le fondateur de l'école française d'analyse des données.

L'idée que, les données devenant de plus en plus massives et les ordinateurs de plus en plus puissants, on arrive a faire l'économie de théorie pour faire de la science prend diverses formes.

L'interprétation des résultats obtenus par des techniques de croisements systématiques, qu'ils soient peu nombreux ou gigantesques avec peu ou beacoup de données, est toujours aussi difficile et ne peut pas s'apuuyer sur les seules techniques. Les mises en garde contre la confusion entre corrélation et cause et effet restent entièrement pertinentes. Le livre de Stephen Jay Gould "la Mal-Mesure de l'Homme" auquel j'ai consacré un exposé, regorge d'exemples de corrélations fortuites et non explicables, le prix du beurre et la distance entre les galaxies.