

## Quelques exemples d'analyse factorielle des correspondances

A.B. Dufour & D. Clot

---

### 1 Responsabilité et Cigarettes

La table de contingence étudiée est extraite du livre de M. Greenacre (1984, [1]). Elle contient la position de l'individu dans l'entreprise et sa relation à la cigarette.

```
entfum <- read.table("http://pbil.univ-lyon1.fr/R/donnees/entfum.txt",h=T,row.names=1)
entfum
```

	non_fumeur	fumeur_occasionnel	fumeur_moyen	gros_fumeur
cadre_experimente	4	2	3	2
cadre_debutant	4	3	7	4
employe_experimente	25	10	12	4
employe_debutant	18	24	33	13
secretaires	10	6	7	2

A l'aide des résultats associés à l'analyse factorielle des correspondances de cette table de contingence, répondre aux questions suivantes.

1. Combien d'axes conserve-t-on ? Justifier.
2. Représenter simultanément les individus et leur addiction à la cigarette. Commenter.
3. La répartition des fumeurs au niveau national est :

```
moynat <- c(42,29,20,9)
moynat <- data.frame(t(moynat))
colnames(moynat) <- colnames(entfum)
moynat
```

	non_fumeur	fumeur_occasionnel	fumeur_moyen	gros_fumeur
1	42	29	20	9

A l'aide des fonctions `suprow` et `points`, placer la moyenne nationale sur la représentation graphique associée aux statuts des individus dans l'entreprise. Commenter.

## 2 Qualité de nouveaux produits

Le problème étudié ici a été soulevé par M. Volle (1989, [2], pages 166-168). Une étude est réalisée sur le lancement d'un nouveau produit de vaisselle. Des consommateurs-testeurs ont essayé huit produits et ont attribué des notes portant sur dix critères liés aux qualités de nettoyage (1), de rinçage (2), de brillance (3), de non rayure (4), à la résistance du produit à l'usage (5), à la conservation des odeurs (6), au fait d'abimer ou non les ongles (7), à la forme du contenant permettant d'aller ou non dans les coins (8), de bien récurer (9), d'être bien tenu en main (10). Les notes sont définies à l'aide d'une échelle bipolaire comprenant six états possibles comme explicité dans l'exemple ci-dessous :

raye			X				ne raye pas
------	--	--	---	--	--	--	-------------

Le résultat peut être considéré de deux façons différentes :

**Cas A** : comme une seule note sur 6 points : 2/6 pour le critère global 'rayure',

**Cas B** : comme deux notes sur 6 mesurant à la fois l'intensité de la qualité et du défaut du critère étudié : 2 pour 'le produit raye' et 4 pour 'le produit ne raye pas'.

Dans les deux cas, les notes des consommateurs sont sommées les unes aux autres.

### Cas A : une note globale

Les données se présentent sous la forme du tableau ci-dessous :

```
casA <- read.table("http://pbil.univ-lyon1.fr/R/donnees/qualprodA.txt", h=TRUE, row.names=1)
casA
      P1 P2 P3 P4 P5 P8 P9 P10
nettoyage 364 640 507 629 529 406 680 697
resistance 521 544 185 502 577 376 620 660
rincage    325 584 298 596 390 396 646 651
odeur      464 648 407 554 455 452 624 632
ongles     463 432 365 619 143 416 486 530
coins      234 588 585 568 450 461 604 659
tenu       637 560 576 708 319 544 546 703
rayure     469 456 583 579 135 319 523 504
brillance  107 415 626 378 161 320 433 431
recurage   389 670 477 569 658 433 694 717
```

Ce tableau pourrait être considéré comme une table de contingence car la somme d'une ligne et la somme d'une colonne ont un sens. La somme des colonnes nous donne les points attribués à tel ou tel produit. Plus la somme est élevée, plus le produit est de qualité. La somme des lignes nous donne les points attribués à tel ou tel critère.

```
apply(casA,1,sum)
nettoyage resistance rincage odeur ongles coins tenu
      4452      3985      3886      4236      3454      4149      4593
      rayure brilliance recurage
      3568      2871      4607

apply(casA,2,sum)
      P1 P2 P3 P4 P5 P8 P9 P10
3973 5537 4609 5702 3817 4123 5856 6184
```

1. Réaliser une AFC de ce tableau et interpréter les résultats.
2. Noter les positions des produits et des critères de qualité en fonction des marges. Quel phénomène observe-t-on ?

### Cas B : une note pour la qualité et une note pour le défaut

Les données se présentent sous la forme du tableau ci-dessous :

```
casB <- read.table("http://pbil.univ-lyon1.fr/R/donnees/qualprodB.txt", h=TRUE, row.names=1)
casB
      P1 P2 P3 P4 P5 P8 P9 P10
nettoie_bien 364 640 507 629 529 406 680 697
nettoie_mal 404 128 261 139 239 362 88 71
resistant 521 544 185 502 577 376 620 660
non_resistant 247 224 583 266 191 392 148 108
rincer_facile 325 584 298 596 390 396 646 651
rincer_difficile 443 184 470 172 378 372 122 117
odeurs_enlevees 464 648 407 554 455 452 624 632
odeurs_conservees 304 120 361 214 313 316 144 136
ongles_saufs 463 432 365 619 143 416 486 530
ongles_abimes 305 336 403 149 625 352 282 238
coins_oui 234 588 585 568 450 461 604 659
coins_non 534 180 183 200 318 307 164 109
tenu_main 637 560 576 708 319 544 546 703
nontenu_main 131 208 192 60 449 224 222 65
raye_non 469 456 583 579 135 319 523 504
raye_oui 299 312 185 189 633 449 245 264
brillance 107 415 626 378 161 320 433 431
non_brillance 661 353 142 390 607 448 335 337
recure_bien 389 670 477 569 658 433 694 717
recure_mal 379 98 291 199 110 335 74 51
```

Ce nouveau fichier présente une caractéristique particulière. La somme des notes associées à un critère de jugement est constante.

```
margeL <- apply(casB,1,sum)
tri <- seq(1,10,by=2)
sapply(tri, function(x) margeL[x]+margeL[x+1])
      nettoie_bien      resistant      rincer_facile      odeurs_enlevees      ongles_saufs
      6144              6144              6144              6144              6144
```

Réaliser une analyse des correspondances sur ce tableau en étudiant plus finement :

- a) la position des critères de qualité et de défaut,
- b) la position des produits 1, 3, 5 et 10.

## 3 Crimes aux USA de 1965 à 2005

Les données sont réparties dans quatre colonnes : l'état, l'année de recueil de l'information, le nombre d'habitants et le nombre de crimes violents.

```
[1] "Etat"           "Date"           "Population"     "Crime_Violent"
```

La table de contingence associant les états et les années (les dates posées en variable qualitative) montre que l'état de New-York n'a pas été échantillonné de 1960 à 1964. C'est pourquoi on étudiera l'évolution de la criminalité de 1965 à 2005. Les années sont passées en colonnes et les états en lignes.

```
annees <- as.character(1965:2005)
ans90 <- as.character(65:99)
ans90 <- paste("A",ans90,sep="")
ans2000 <- as.character(0:5)
ans2000 <- paste("A0",ans2000,sep="")
ans <- c(ans90,ans2000)
#
crimes <- matrix(0,ncol=length(annees),nrow=51)
for (i in 1:length(annees)) crimes[,i] <- CSD$Crime_Violent[CSD$Date==annees[i]]
crimes <- as.data.frame(crimes)
rownames(crimes) <- unique(CSD$Etat)
colnames(crimes) <- ans
```

`crimes` est une table de contingence du nombre de crimes violents commis dans les 51 Etats américains entre 1965 et 2005.

1. Réaliser une analyse factorielle des correspondances sur cette table de contingence. Expliquer le choix du nombre d'axes conservés.
2. Interpréter les modalités liées aux années.
3. Discuter de la position des pays suivants : Connecticut (ligne 7), District of Columbia (ligne 9), Nevada (ligne 29) et New-York (ligne 33).
4. Le Colorado (ligne 6), le Montana (ligne 27) et le Vermont (ligne 46) sont-ils bien représentés sur le premier plan factoriel? Si oui, interpréter également leur position. Si non, rechercher le ou les axes mettant en évidence ces Etats et interpréter.
5. On pourrait également réaliser une analyse en composantes principales du tableau `crimes`. A quelle question répond-on quand on réalise une telle analyse?
6. A quelle question répond-on quand on réalise une analyse factorielle des correspondances?

## Références

- [1] M.J. Greenacre. *Theory and Applications of correspondence analysis*. Academic Press, London, 1984.
- [2] M. Volle. *Analyse des donn.* Economica, Paris, economica edition, 1981.