Consultation statistique avec le logiciel 

Comment retrouver une analyse des correspondances intra-classes dans une analyse des correspondances internes ?

D. Chessel

14 mars 2006

Gaël Didier s'étonne de ne pas retrouver dans une analyse des correspondances internes les résultats obtenus par une analyse des correspondances intra-classes qui est un cas particulier. La fiche explique pourquoi. Il fait les calculs dans ADE-4. En voulant les reproduire dans ade4, on met en évidence une erreur signalée et corrigée dans `witwit.coa` par Campo Elías PARDO.

Table des matières

1 Introduction

La discussion porte sur les tableaux édités dans [1, page **8**-170] dont on disposera par :

```
x <- read.table("http://pbil.univ-lyon1.fr/R/donnees/scolar.txt",
    header = TRUE, as.is = TRUE)
x
```

```
   sexe diplo  agri inge tech ouvqua   onq cadsup cadmoy empqua   enq
1   hom  sans 15068    0  302  10143 59394    596   2142   5445  4879
2   hom  bepc  2701  337 1697   3702  8087    296   2801   7348  4987
3   hom   bep  5709  309 2242  30926 17862    892    672   4719  1514
4   hom  bacG   297  917 1969    314  2887   1227   6495   4353  3478
5   hom  bacT  1242    0 1399   1861  1696    298    924   1280   886
6   hom  deug     0  308  367      0     0   2362   2807    614  1326
7   hom   dut   322    0 1943      0     0    318   2301    982     0
8   hom   sup     0 4383  381    337   323   6781   4030      0   661
9   fem  sans  5089    0  281   7470 29997      0   1577  21616 19849
10  fem  bepc  1212    0    0   1859  4334      0   1806  19915  7325
11  fem   bep  1166    0  320   4017  4538      0   4549  32452  6484
12  fem  bacG     0  316  320   1752  1882   2236  17063  16137  5111
13  fem  bacT     0    0  283    657     0    595    875   5865   898
14  fem  deug     0    0    0      0     0    911   4152   1256   294
15  fem   dut     0  304  683    285     0    569  15731   3332   635
16  fem   sup     0 1033    0      0     0   6788   3991   1286     0
```

Il s'agit en fait de deux tableaux séparés sur les élèves scolarisés en 1972-1973, sortis du système éducatif en 1973 et ayant trouvé un emploi. Chaque personne donne le type d'emploi occupé et le niveau de diplôme obtenu. Les codes sont :

`sans` = Sans diplôme obtenu	`bepc` = BEPC
`bep` = BEP ou CAP	`bacG` = Baccalauréat général
`bacT` = Baccalauréat technique	`deug` = DEUG ou ENT
`dut` = DUT ou BTS ou Santé	`sup` = Diplôme de fin d'études supérieures

`agri` = Agriculteurs	`inge` = Ingénieurs
`tech` = Techniciens	`ouvqua` = Ouvriers qualifiés
`onq` = Ouvriers non qualifiés	`cadsup` = Cadres supérieurs
`cadmoy` = Cadres moyens	`empqua` = Employés qualifiés
`enq` = Employés non qualifiés	

Le premier concerne les élèves de sexe masculin :

	agri	inge	tech	ouvqua	onq	csup	cmoy	empqua	enq
sans	15068	0	302	10143	59394	596	2142	5445	4879
bepc	2701	337	1697	3702	8087	296	2801	7348	4987
bep	5709	309	2242	30926	17862	892	672	4719	1514
bacG	297	917	1969	314	2887	1227	6495	4353	3478
bacT	1242	0	1399	1861	1696	298	924	1280	886
deug	0	308	367	0	0	2362	2807	614	1326
dut	322	0	1943	0	0	318	2301	982	0
sup	0	4383	381	337	323	6781	4030	0	661

Le second concerne les élèves de sexe féminin :

	agri	inge	tech	ouvqua	onq	csup	cmoy	empqua	enq
sans	5089	0	281	7470	29997	0	1577	21616	19849
bepc	1212	0	0	1859	4334	0	1806	19915	7325
bep	1166	0	320	4017	4538	0	4549	32452	6484
bacG	0	316	320	1752	1882	2236	17063	16137	5111
bacT	0	0	283	657	0	595	875	5865	898
deug	0	0	0	0	0	911	4152	1256	294
dut	0	304	683	285	0	569	15731	3332	635
sup	0	1033	0	0	0	6788	3991	1286	0

Ces données illustrent la comparaison de tableaux de fréquence binaire dans la terminologie des auteurs : on a affaire à deux tables de contingence ayant les mêmes lignes et les mêmes colonnes. Il s'agit de données très particulières. Les tableaux sont simplement transposé par rapport à la publication d'origine.

2 La question

Gaël Didier pose cette question sur le forum adelist :

```
Bonjour à tous,

J'explique plus en détail mon problème.
Les données forment un tableau de contingence croisant les CSP en
ligne (9 modalités) et les diplômes en colonne (8 modalités)
1 tableau pour les hommes et 1 pour les femmes
(Analyses factorielles simples et multiples, p. 199, Escofier/Pagès).

Le but : faire l'analyse intra-classe, pour supprimer l'effet diplôme.
```

Nous nous plaçons dans les conditions décrites.

```
sexe <- factor(x$sexe)
diplo <- factor(x$diplo)
row.names(x) <- paste(substr(x$sexe, 1, 1), x$diplo, sep = "_")
y <- x[, -c(1, 2)]
y
```

```
        agri inge tech ouvqua   onq cadsup cadmoy empqua   enq
h_sans 15068    0  302  10143 59394    596   2142   5445  4879
h_bepc  2701  337 1697   3702  8087    296   2801   7348  4987
h_bep   5709  309 2242  30926 17862    892    672   4719  1514
h_bacG   297  917 1969    314  2887   1227   6495   4353  3478
h_bacT  1242    0 1399   1861  1696    298    924   1280   886
h_deug     0  308  367      0     0   2362   2807    614  1326
h_dut    322    0 1943      0     0    318   2301    982     0
h_sup      0 4383  381    337   323   6781   4030      0   661
f_sans  5089    0  281   7470 29997      0   1577  21616 19849
f_bepc  1212    0    0   1859  4334      0   1806  19915  7325
f_bep   1166    0  320   4017  4538      0   4549  32452  6484
f_bacG     0  316  320   1752  1882   2236  17063  16137  5111
f_bacT     0    0  283    657     0    595    875   5865   898
f_deug     0    0    0      0     0    911   4152   1256   294
f_dut      0  304  683    285     0    569  15731   3332   635
f_sup      0 1033    0      0     0   6788   3991   1286     0
```

Nous sommes à pied d'œuvre. La question concerne la comparaison entre une analyse des correspondances intra-classes de procédure :

1. `COA : COrrespondence Analysis`

2. `Discrimin : Initialize : LinkPrep`
3. `Discrimin : Within Analysis : Run`

et l'analyse des correspondances internes `COA : Internal COA` dans ADE-4. Dans ade4, les fonctions correspondantes sont `coa`, `within` et `witwit.coa`. Il est intéressant de tester l'idée qu'une analyse des correspondances internes étant une double analyse des correspondances intra, on doit retrouver les résultats d'une analyse des correspondances intra classes avec une analyse des correspondances internes. Je vérifie que c'est vrai dans ade4 et je cherche à savoir où est la question pour Gaël Didier. Il dit :

```
Pour ce faire on prend le tableau juxtaposé
horizontalement (homme, femme), que l'on transpose, puisque
dans within analysis: run de Discrimin, ce sont les
lignes qui sont partitionnées.

*Analyse intra via Discrimin*
Procédure :
1/ AFC (COA) du tableau initial transposé --> triplet
statistique ;
2/ Discrimin : Initialize LinkPrep :
     le triplet est celui issu de l'AFC ;
     categories file : déterminé à partir des libellés des
lignes :diplôme 1 ... diplôme 8 diplôme 1 diplôme 8;
On retrouve bien les bonnes valeurs (programmation manuelle)
d'inertie inter et intra-classes : 0.709% et 0.291%.
3/ Discrimin : within analysis: run , en utilisant le
fichier de l'étape 2.

Résultats
Valeurs propres  \% inertie  \% cumulé
01   +2.3475E-01 +0.6085 +0.6085
02   +5.9647E-02 +0.1546 +0.7631
03   +4.0317E-02 +0.1045 +0.8676
04   +3.4937E-02 +0.0906 +0.9581
05   +8.4285E-03 +0.0218 +0.9800
06   +6.7795E-03 +0.0176 +0.9975
07   +5.9281E-04 +0.0015 +0.9991
08   +3.5278E-04 +0.0009 +1.0000
09   +0.0000E+00 +0.0000 +1.0000

Je me contente de donner les valeurs propres. Le reste en découle.
```

Pour refaire ceci :

```
options(digits = 4)
library(ade4)
coa1 <- dudi.coa(y, scan = F)
w1 <- within(coa1, diplo, scan = F)
w1$eig
```

```
[1] 0.2347538 0.0596465 0.0403167 0.0349373 0.0084285 0.0067795 0.0005928 0.0003528
```

```
cumsum(w1$eig/sum(w1$eig))
```

```
[1] 0.6085 0.7631 0.8676 0.9581 0.9800 0.9975 0.9991 1.0000
```

Nous sommes bien d'accord. Il continue :

```
Puisque l'analyse interne est équivalente à une analyse
intra lorsque les colonnes ne sont pas partitionnées, on en
réalise une:

*Analyse Interne via COA*

Fichier d'entrée : le fichier initial transposé ;
Row indicator : les lignes sont partitionnées en 8/8 (8
lignes hommes puis 8 lignes femmes) ;
Col indicator : par défaut, puisque les colonnes ne sont pas
partitionnées ;

Résultats
Valeurs propres, % inertie; % cumulé
01   +5.3622E-01 +0.4871 +0.4871
02   +2.0407E-01 +0.1854 +0.6725
03   +1.4025E-01 +0.1274 +0.7999
04   +1.1485E-01 +0.1043 +0.9042
05   +4.5884E-02 +0.0417 +0.9459
06   +3.8065E-02 +0.0346 +0.9805
07   +1.9988E-02 +0.0182 +0.9986
08   +1.5002E-03 +0.0014 +1.0000
09   +0.0000E+00 +0.0000 +1.0000

Les valeurs propres entre les 2 méthodes sont totalement
différentes.

Celles de la méthode Discrimin correspondent à celles
données par Escofier/Pagès.

Les valeurs propres de la méthode Internal COA sont
identiques à la programmation que j'ai faite selon l'article
de Cazes/Moreau : Analyse des correspondances d'un tableau
de contingence dont les lignes et les colonnes sont munies
d'une structure de graphe bistochastique (dans  l'Analyse
des correspondances et les techniques connexes) :
Analyse de Benzécri, 1983 :
AFC sur le tableau
k_{ij}-k_{i.}*k_{pj}/k_{p.}+k_{i.}*k_{.j}/k, où k_{ij} est
le tableau initial, les lignes sont partitionnées selon p
classes, k est la comme des k_{ij}, les autres notations
étant assez évidentes je pense.
```

3 Une jolie erreur

Pour retrouver immédiatement ce que trouve l'auteur :

```
w2 <- within(coa1, sexe, scan = F)
w2$eig
```

```
[1] 0.536219 0.204068 0.140251 0.114854 0.045884 0.038065 0.019988 0.001500
```

```
cumsum(w2$eig/sum(w2$eig))
```

```
[1] 0.4871 0.6725 0.7999 0.9042 0.9459 0.9805 0.9986 1.0000
```

L'analyse intra classe est définie par un facteur (sexe) qui a deux modalités (comme chacun sait) et donc définit deux blocs. En faisant l'analyse interne avec deux blocs (8 et 8), on utilise bien sûr le facteur sexe de l'intra. Donc l'analyse interne (8 et 8) est bien l'analyse intra sexe et pas du tout l'analyse intra diplôme. Ce qui est bien cohérent :

```
w3 <- witwit.coa(coa1, c(8, 8), 9, scan = F)
w3$eig
```

```
[1] 0.536219 0.204068 0.140251 0.114854 0.045884 0.038065 0.019988 0.001500
```

```
cumsum(w3$eig/sum(w3$eig))
```

```
[1] 0.4871 0.6725 0.7999 0.9042 0.9459 0.9805 0.9986 1.0000
```

4 Une erreur bien utile

Mais cette erreur est bien utile, parce que la fonction witwit.coa ne reproduisait pas exactement celle de ADE-4. Pour retrouver l'identité entre les deux, je me suis souvenu d'avoir reçu un mail de Campo Elías PARDO, auquel je n'avais pas répondu :

```
I'm sending again a question about function witwit.coa

In internal correspondence analysis, I wanted to compute the desviations
between the  frequency table and the model used in ICA,  through:
ica <- witwit.coa(.........)
des <- ica$tab * ica$lw %*% t(ica$cw)
I hope des = 0 but it is different
I look for the reason  of this result. I found  the causa:
    tabinit <- data.frame(tabinit+wrmat)
why is wrmat summed? Is this an error? The correct sentence is:
    tabinit <- data.frame(tabinit)
Is there any intention in the sum  of wrmat?
Thanks,
```

Oui, c'était une erreur et je remercie l'auteur. La fonction sera mise à jour dès que possible. Ceci permet de retrouver les résultats de départ.

```
args(witwit.coa)
```

```
function (dudi, row.blocks, col.blocks, scannf = TRUE, nf = 2)
NULL
```

```
z <- y[c(1, 9, 2, 10, 3, 11, 4, 12, 5, 13, 6, 14, 7, 15, 8, 16),
    ]
coa2 <- dudi.coa(z, scan = F)
w4 <- witwit.coa(coa2, rep(2, 8), 9, scan = F)
w4$eig
```

```
[1] 0.2347538 0.0596465 0.0403167 0.0349373 0.0084285 0.0067795 0.0005928 0.0003528
```

```
cumsum(w4$eig/sum(w4$eig))
```

```
[1] 0.6085 0.7631 0.8676 0.9581 0.9800 0.9975 0.9991 1.0000
```

La boucle est bouclée. Gaël Didier n'a pas associé la définition des blocs à la bonne variable. Il a fait une erreur et moi j'en fait deux. La première est dans une ligne de la fonction. La seconde est de ne pas avoir tenu compte du signalement de Campo Elías PARDO. Les sources ouverts, ça change bien des choses. Merci à tous les deux. On y a gagné un bug de moins et un bon exemple de plus.

Références

[1] B. Escofier and J. Pagès. *Analyses factorielles simples et multiples : objectifs, méthodes et interprétation*. Dunod, Paris, 1990.