## 4.5 Etude plus précise de l'ACM sur un mini-exemple

### 4.5.1 L'exemple choisi

L'exemple qui suit est tiré de [Rouanet - Le Roux] qui fait lui-même référence à une célèbre enquête britannique (D.V. Glass, 1954, Social Mobility in Britain, London, Routledge & Kegan Paul).

Une enquête a été menée auprès de 3450 individus. Les variables qui ont été observées sont les suivantes :

- A : statut du père du répondant (deux modalités : a1 élevé, a2 faible)
- B : niveau scolaire (deux modalités : b1 élevé, b2 faible)
- C : statut du répondant (trois modalités : c1, c2, c3 de niveaux décroissants.

Le tableau des effectifs est donné par :

| A | B | C | $n_k$ |
|---|---|---|---|
| a1 | b1 | c1 | 106 |
| a1 | b1 | c2 | 88 |
| a1 | b1 | c3 | 8 |
| a1 | b2 | c1 | 18 |
| a1 | b2 | c2 | 45 |
| a1 | b2 | c3 | 11 |
| a2 | b1 | c1 | 106 |
| a2 | b1 | c2 | 776 |
| a2 | b1 | c3 | 133 |
| a2 | b2 | c1 | 27 |
| a2 | b2 | c2 | 1274 |
| a2 | b2 | c3 | 858 |
| TOTAL | | | 3450 |

Nous avons en tout 7 modalités, et 2x2x3 = 12 patrons de réponses possibles. Tous sont représentés parmi les réponses observées. Le début du tableau disjonctif complet est le suivant :

| | **a1** | **a2** | **b1** | **b2** | **c1** | **c2** | **c3** |
|---|---|---|---|---|---|---|---|
| sujet 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| sujet 2 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| ... | ... | 0 | ... | ... | ... | ... | ... |

Les données observées peuvent également être décrites à l'aide du *tableau disjonctif des patrons* :

| | **a1** | **a2** | **b1** | **b2** | **c1** | **c2** | **c3** | **TOTAL** |
|---|---|---|---|---|---|---|---|---|
| a1b1c1 | 106 | 0 | 106 | 0 | 106 | 0 | 0 | **318** |
| a1b1c2 | 88 | 0 | 88 | 0 | 0 | 88 | 0 | **264** |
| a1b1c3 | 8 | 0 | 8 | 0 | 0 | 0 | 8 | **24** |
| a1b2c1 | 18 | 0 | 0 | 18 | 18 | 0 | 0 | **54** |
| a1b2c2 | 45 | 0 | 0 | 45 | 0 | 45 | 0 | **135** |
| a1b2c3 | 11 | 0 | 0 | 11 | 0 | 0 | 11 | **33** |
| a2b1c1 | 0 | 106 | 106 | 0 | 106 | 0 | 0 | **318** |
| a2b1c2 | 0 | 776 | 776 | 0 | 0 | 776 | 0 | **2328** |
| a2b1c3 | 0 | 133 | 133 | 0 | 0 | 0 | 133 | **399** |
| a2b2c1 | 0 | 27 | 0 | 27 | 27 | 0 | 0 | **81** |
| a2b2c2 | 0 | 1274 | 0 | 1274 | 0 | 1274 | 0 | **3822** |
| a2b2c3 | 0 | 858 | 0 | 858 | 0 | 0 | 858 | **2574** |
| **TOTAL** | **276** | **3174** | **1217** | **2233** | **257** | **2183** | **1010** | **10350** |

Enfin, le tableau de Burt est ici :

| | A:a1 | A:a2 | B:b1 | B:b2 | C:c1 | C:c2 | C:c3 | **Total** |
|---|---|---|---|---|---|---|---|---|
| A:a1 | 276 | 0 | 202 | 74 | 124 | 133 | 19 | **828** |
| A:a2 | 0 | 3174 | 1015 | 2159 | 133 | 2050 | 991 | **9522** |
| B:b1 | 202 | 1015 | 1217 | 0 | 212 | 864 | 141 | **3651** |
| B:b2 | 74 | 2159 | 0 | 2233 | 45 | 1319 | 869 | **6699** |
| C:c1 | 124 | 133 | 212 | 45 | 257 | 0 | 0 | **771** |
| C:c2 | 133 | 2050 | 864 | 1319 | 0 | 2183 | 0 | **6549** |
| C:c3 | 19 | 991 | 141 | 869 | 0 | 0 | 1010 | **3030** |
| **Total** | **828** | **9522** | **3651** | **6699** | **771** | **6549** | **3030** | **31050** |

### 4.5.2 Lien entre l'ACM et l'AFC

Comme l'indiquent Rouanet et Le Roux :

*Effectuer l'analyse des correspondances multiples, c'est effectuer l'analyse factorielle des correspondances du tableau disjonctif complet, muni des relations K<Q> (modalités emboîtées dans les questions) et I<K<q>> (individus emboîtés dans les modalités de chaque question).*

Nous pouvons donc, comme en AFC, nous intéresser aux profils ligne et colonne, aux taux de liaison et au $\Phi^2$ du tableau disjonctif complet, vu comme un tableau de contingence. Mais, ce tableau comporte 3450 lignes ! Cependant, nous avons vu que la métrique du $\Phi^2$, utilisée pour l'AFC, possède la propriété d'équivalence distributionnelle : si on regroupe deux lignes correspondant au même patron de réponses, on ne change rien aux autres profils lignes, ni aux autres profils colonnes. Autrement dit, on retrouvera les mêmes résultats en effectuant une AFC sur le tableau disjonctif des patrons.

Comme en AFC, on peut calculer des fréquences, des fréquences lignes, des fréquences colonnes et des profils lignes et profils colonnes moyens. Affichez la feuille "fréquences" du classeur Excel Statut.xls et observez les fréquences ainsi obtenues.

L'élément le plus facile à interpréter est le profil colonne moyen : ce sont les fréquences des différents patrons de réponses dans la population étudiée.

#### 4.5.2.1 Distances entre profils lignes

En AFC, nous avions donné les formules permettant de calculer les distances entre deux profils lignes ou entre deux profils colonnes. La distance utilisée est la *métrique du* $\Phi^2$. Ici, compte tenu de la structure particulière du tableau de contingence utilisé, les formules indiquées deviennent :

$$d^2_{\Phi^2}(L_i, L_{i'}) = \frac{1}{Q}\sum_k \frac{(\delta_{ik} - \delta_{i'k})^2}{f_{\bullet k}}$$

Notations utilisées : Li et Li' désignent deux patrons, Q est le nombre de questions. $\delta_{ik}$ prend la valeur 1 si la modalité k fait partie du patron i, et la valeur 0 sinon. Enfin, $f_{\bullet k}$ est la fréquence de la modalité k dans la population.

Par exemple, sachant que les fréquences des modalités b1, b2, c1 et c3 sont respectivement : 35,27%, 64,72%, 7,45% et 29,28%, la distance entre les deux patrons a1b1c1 et a1b2c3 est donnée par :

$$d^2_{\Phi^2}(a1b1c1, a1b2c3) = \frac{1}{3}\left(\frac{1}{0,3527} + \frac{1}{0,6472} + \frac{1}{0,0745} + \frac{1}{0,2928}\right) = 7,07$$

Autrement dit, deux individus (ou deux patrons) sont d'autant plus éloignés que leurs réponses diffèrent pour un plus grand nombre de questions et pour des modalités rares.

La distance d'un patron au profil ligne moyen est :

$$d^2_{\Phi^2}(O,\ L_i) = \left(\frac{1}{Q}\sum_k \frac{\delta_{ik}}{f_{\bullet k}}\right) - 1$$

Par exemple, sachant que les modalités a1, b1 et c1 ont pour fréquences respectives : 8,00%, 35,27% et 7,45%, la distance du patron a1b1c1 à l'origine est :

$$d^2_{\Phi^2}(O,\ a1b1c1) = \frac{1}{3}\left(\frac{1}{0,08} + \frac{1}{0,3527} + \frac{1}{0,0745}\right) - 1 = 8,59$$

Autrement dit, un patron sera d'autant plus loin de l'origine qu'il fait intervenir des modalités plus rares.

La contribution (absolue) d'un patron à la variance du nuage est obtenue en multipliant la distance précédente par la fréquence du patron dans la population. Pour a1b1c1 :

$$Cta(a1b1c1) = \frac{106}{3450} \times 8,59 = 0,2638$$

Sa contribution relative est obtenue en divisant par la variance totale du nuage (cf plus bas), c'est à dire 1,33. Autrement dit :

$$Ctr(a1b1c1) = \frac{0,2638}{1,33} = 0,1979$$

### 4.5.2.2 Distances entre profils colonnes

La distance entre les modalités k et k' est donnée par :

$$d^2_{\Phi^2}(M_k,\ M_{k'}) = \frac{1}{f_{\bullet k}} + \frac{1}{f_{\bullet k'}} - 2\frac{f_{kk'}}{f_{\bullet k} f_{\bullet k'}}$$

où $f_{kk'}$ est la fréquence de la combinaison de modalités k et k'.

Par exemple, sachant que les fréquences de a1 et b1 sont 8,00% et 35,27%, et que la fréquence de la combinaison a1b1 est 5,86%, on obtient :

$$d^2_{\Phi^2}(a1,\ b1) = \frac{1}{0,08} + \frac{1}{0,3527} - 2\frac{0,0586}{0,08 \times 0,3527} = 11,18$$

La distance d'une modalité au profil colonne moyen est donnée par :

$$d^2_{\Phi^2}(O,\ M_k) = \frac{1}{f_{\bullet k}} - 1$$

Pour la modalité a1, on obtient, par exemple :

$$d^2_{\Phi^2}(O,\ a1) = \frac{1}{0,08} - 1 = 11,5$$

Autrement dit, une modalité sera d'autant plus loin du profil moyen que sa fréquence est faible.

La contribution absolue d'une modalité à la variance du nuage de points est :

$$Cta(M_k) = \frac{1 - f_{\bullet k}}{Q}$$

Par exemple, pour la modalité a1 :

$$Cta(a1) = \frac{1 - 0,08}{3} = 0,3067$$

Sa contribution relative est obtenue en divisant par la variance totale du nuage (1,33 dans notre exemple) :

$$Ctr(a1) = \frac{0,3067}{1,33} = 0,23$$

### 4.5.2.3 Taux de liaison et Phi-2

Pour le tableau disjonctif complet, ou le tableau disjonctif des patrons, considérés comme des tableaux de contingence, le coefficient Phi-2 vaut :

$$\Phi^2 = \frac{K}{Q} - 1$$

où K désigne le nombre de modalités et Q le nombre de questions

Dans notre exemple, on a : K=7, Q=3, et donc : $\Phi^2 = \frac{7}{3} - 1 = 1,33.$

## 4.5.3 Valeurs propres

Les fichiers contenus dans le répertoire Statut du serveur de TD permettent de retrouver les résultats qui suivent.

Chargez Statistica et ouvrez la feuille de données statut-disjonctif-patrons.sta.
Exécutez ensuite une AFC en indiquant que la feuille de données est un tableau de contingence.
Vous devriez obtenir le tableau des valeurs propres (non nulles) suivant :

| Valeurs Propres et Inertie de toutes les Dimensions (statut-disjonctif-patrons.sta) | | | | | |
|---|---|---|---|---|---|
| Table d'Entrée (Lignes x Colonnes) : 12 x 7 | | | | | |
| Inertie Totale = 1,3333 Chi2 = 13800, dl = 66 p = 0,0000 | | | | | |
| | ValSing. | ValProp. | %age | %age | Chi_ |
| | | | Inertie | Cumulé | |
| 1 | 0,7466 | 0,5574 | 41,81 | 41,81 | 5769,17 |
| 2 | 0,5983 | 0,3579 | 26,84 | 68,65 | 3704,29 |
| 3 | 0,4824 | 0,2328 | 17,46 | 86,11 | 2409,25 |
| 4 | 0,4304 | 0,1852 | 13,89 | 100,00 | 1917,27 |

Nous avions constaté sur les exemples précédents que la décroissance des valeurs propres est assez lente. Pour donner une plus juste valeur de l'importance relative des axes principaux, Benzécri a proposé de considérer des taux modifiés, calculés de la manière suivante :

- On ne considère que les valeurs propres $\lambda_l$ supérieures à la valeur moyenne $\frac{1}{Q}$.
- On remplace la valeur propre $\lambda_l$ par $\lambda'_l = \left(\frac{Q}{Q-1}\right)^2 \left(\lambda_l - \frac{1}{Q}\right)^2$.
- On calcule le taux correspondant à $\lambda_l$ en divisant $\lambda'_l$ par la somme des $\lambda'_l$.

Dans notre exemple, deux valeurs propres sont supérieures à la moyenne (1/3). Les valeurs modifiées conduisent aux résultats suivants :

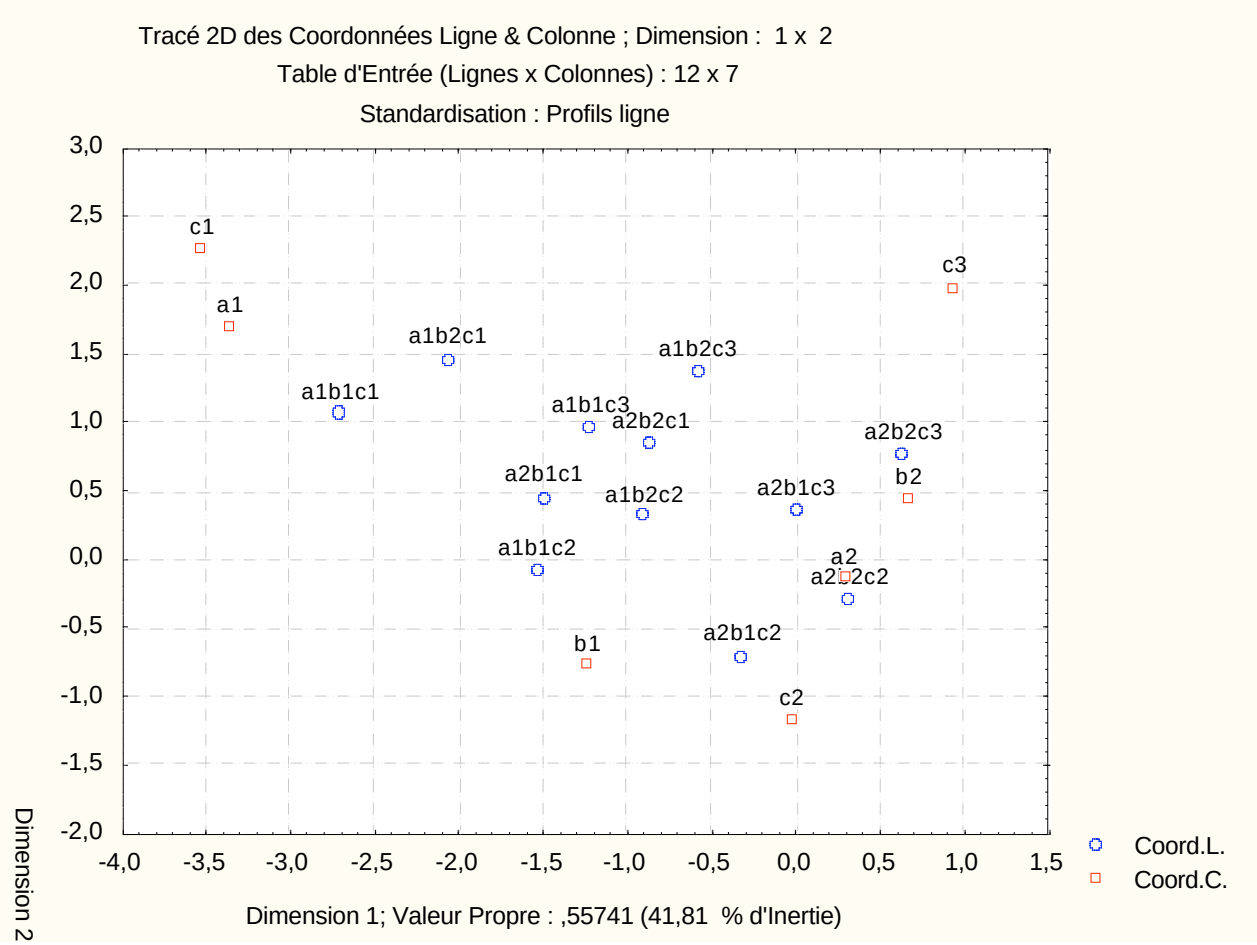
| Valeur propre | $\lambda'_l$ | Taux modifié |
|---|---|---|
| 0,5574 | 0,1130 | 98,81% |
| 0,3579 | 0,0014 | 1,19% |

On pourrait donc se borner à étudier l'axe 1.

### 4.5.4 Résultats graphiques

Les graphiques produits possèdent des propriétés géométriques intéressantes. Cependant, nous avions jusqu'à présent utilisé l'option "Centrer-réduire les données - Profils ligne et colonne" (sous l'onglet "Options" de la fenêtre de dialogue). Or, la mise en évidence de ces propriétés nécessite d'utiliser, selon le cas, l'option "Profils ligne (interpréter dist. lignes)" ou l'option "Profils colonne (interpréter dist. colonnes)".

Par exemple, en utilisant les profils lignes, on obtient :

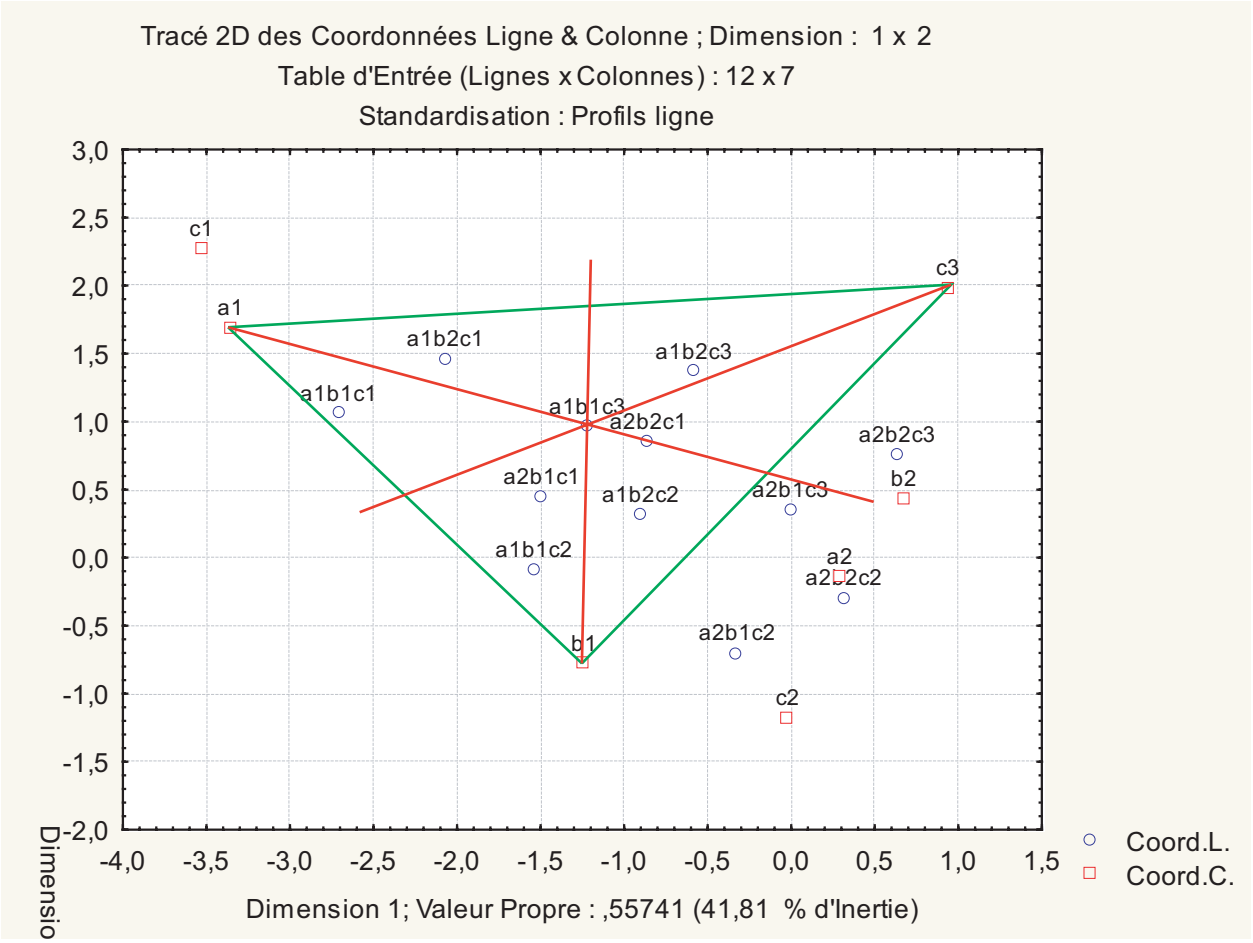



Avec ce choix d'échelles, le nuage des patrons est entièrement contenu à l'intérieur de celui les modalités. Ce graphique met particulièrement bien en évidence les propriétés suivantes :

*Le point représentant chaque patron est l'équibarycentre des modalités correspondant à ce patron.*

Cette propriété est vraie aussi bien pour les individus que pour les patrons. Elle est vraie dans l'espace multidimensionnel, et elle est conservée par les projections sur les plans factoriels.

De manière plus claire, chaque patron est le centre de gravité du triangle formé par ses trois modalités. Ou encore, par exemple, considérons les modalités a1, b1, c3 et le patron a1b1c3. Chaque droite joignant l'une de ces modalités au point représentant le patron passe par le milieu du segment défini par les deux autres modalités :

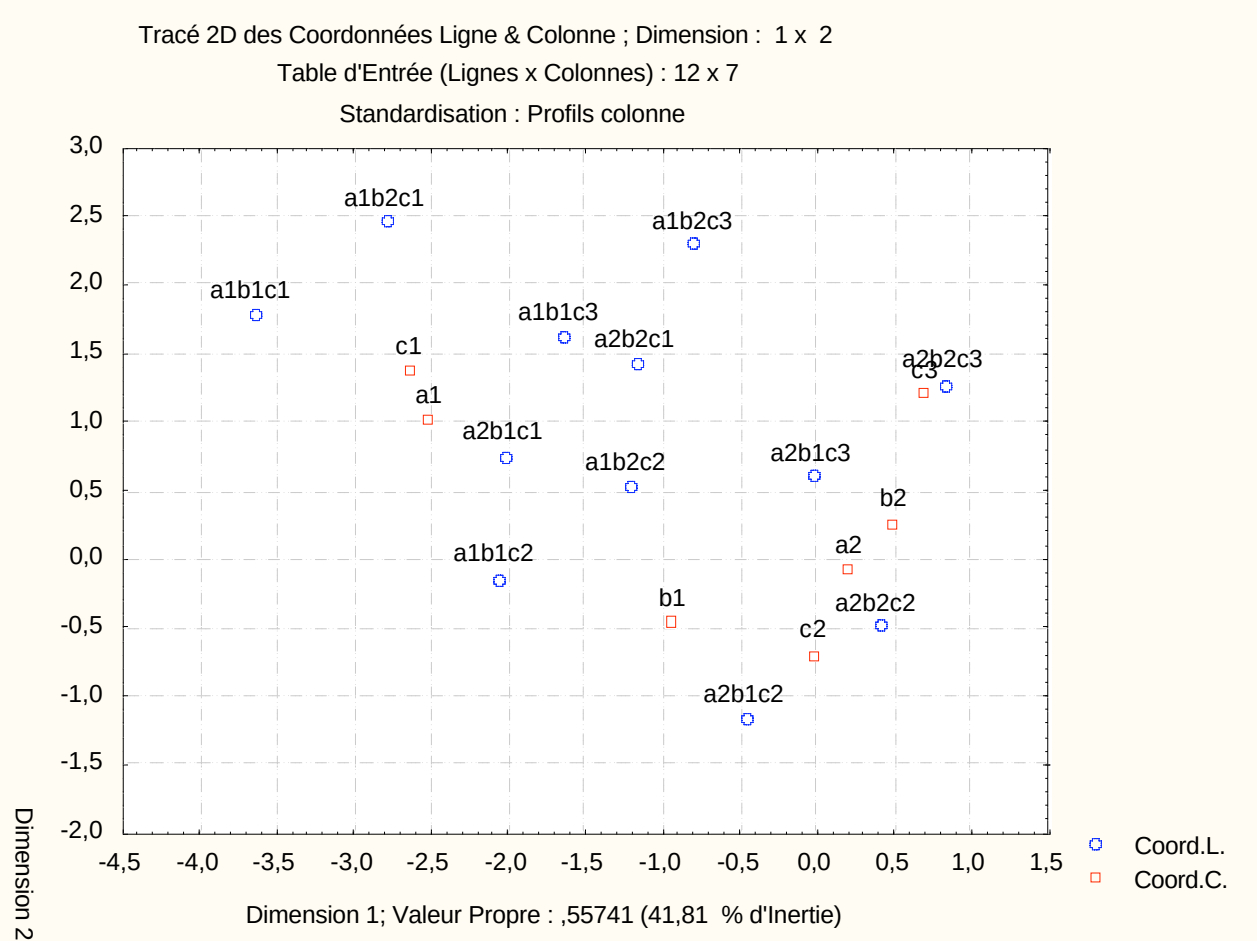



On constate également sur le graphique que les droites (a1 a2) et (b1 b2) passent par l'origine du repère, qui est également à l'intérieur du triangle formé par les trois points c1, c2 et c3.

C'est une conséquence de la propriété suivante :

Le sous-nuage des modalités d'une question a pour point moyen le point moyen du nuage (moyenne pondérée par les fréquences des modalités).

En utilisant les profils colonnes, on obtient :




Chaque modalité est obtenue comme barycentre des patrons qui l'ont choisie, mais chaque patron doit être pondéré par sa fréquence. La constatation la plus immédiate que l'on peut faire sur le graphique ci-dessus est la suivante : chaque modalité se trouve à l'intérieur du polygone (convexe) défini à partir des 4 ou 6 patrons qui l'ont choisie.

On voit également apparaître sur cette représentation la propriété *d'équipollence*, vérifiée par les patrons de réponses. Par exemple les 4 segments suivants :

- le segment qui joint a1b1c1 à a1b1c2
- le segment qui joint a1b2c1 à a1b2c2
- le segment qui joint a2b1c1 à a2b1c2
- le segment qui joint a2b2c1 à a2b2c2

sont parallèles, de même longueur et de même sens

### 4.5.5 L'analyse d'un tableau de Burt

Dans un exposé théorique sur l'ACM, tels que ceux de [Crucianu] ou de [Rouanet, Le Roux], l'analyse du tableau de Burt est distinguée de celle du TDC ou du tableau disjonctif des patrons. Il est notamment indiqué que les valeurs propres produites par cette analyse sont les carrés des valeurs propres précédentes, et que le Phi-2 du tableau de Burt n'est pas celui du TDC. Cependant, les représentations graphiques produites (limitées aux seules modalités) peuvent être interprétées de façon analogue.

Qu'en est-il avec Statistica ?
Ouvrez la feuille de données Statistica Statut-Burt.sta.
Effectuez l'analyse en choisissant l'onglet "Analyse des Correspondances Multiple" et l'item "Tableau de Burt".

On constate que l'on obtient, pour les modalités, des résultats identiques aux précédents. En particulier, les valeurs propres sont celles qui ont indiquées plus haut. Ce sont également celles que l'on obtiendrait en effectuant l'analyse à partir de l'onglet "Analyse des Composantes Multiples" et du tableau protocole ou du tableau des effectifs.

En revanche, nous pouvons effectuer une AFC à l'aide de l'onglet "Analyse des correspondances", en spécifiant le tableau de Burt comme tableau de contingence. On retrouve alors les résultats indiqués dans les exposés théoriques. Par exemple, le tableau des valeurs propres est alors donné par :

| Nombre de Dims. | Valeurs Propres et Inertie de toutes les Dimensions (statut-Burt.sta)<br>Inertie Totale = ,52730 Chi2 = 16373, dl = 36 p = 0,0000 | | | | |
|---|---|---|---|---|---|
| | ValSing. | ValProp. | %age | %age | Chi2 |
| | | | Inertie | Cumulé | |
| 1 | 0,5574 | 0,3107 | 58,9236 | 58,9236 | 9647,3580 |
| 2 | 0,3579 | 0,1281 | 24,2925 | 83,2162 | 3977,3286 |
| 3 | 0,2328 | 0,0542 | 10,2761 | 93,4922 | 1682,4691 |
| 4 | 0,1852 | 0,0343 | 6,5078 | 100,0000 | 1065,4910 |

### 4.5.6 Synthèse : menu à utiliser selon la forme des données d'entrée

On a observé plusieurs (plus de 2) variables nominales sur une population, et on souhaite explorer ces données à l'aide d'une ACM. Selon la forme sous laquelle ces données sont disponibles, on utilisera sous Statistica les menus suivants :

| Format des données | Onglet "Analyse des Correspondances" | Onglet "Analyse des Correspondances Multiple" | Observations |
|---|---|---|---|
| Tableau protocole | Non | Oui | AFC impossible si plus de 2 variables |
| Tableau d'effectifs | Non | Oui | AFC impossible si plus de 2 variables |
| Tableau Disjonctif Complet | Oui | Non | |
| Tableau Disjonctif des patrons | Oui | Non | |
| Tableau de Burt | Oui | Oui | Les deux analyses ne fournissent pas les mêmes résultats |

## *4.6 Exercice*

Les fichiers Beverage.sta et Beverag2.sta sont des fichiers d'exemples fournis avec Statistica pour illustrer l'ACM. Ils ont été recopiés dans le répertoire Boissons du serveur de TD.

Le fichier d'exemple Beverage.sta contient des données collectées sur un groupe d'étudiants en maîtrise de gestion, hommes et femmes, de l'Université de Columbia, auxquels on a demandé d'indiquer la fréquence avec laquelle ils avaient acheté et consommé différents types de soda durant du mois écoulé. Les données pour les 34 individus ont été codifiées dans un tableau disjonctif complet (binaire) : un 1 a été saisi si l'individu a répondu avoir acheté ou consommé au moins une fois au cours du mois la boisson respective, et un 0 a été saisi si l'individu respectif a répondu avoir acheté ou consommé moins d'une fois dans le mois. Pour chacun des 8 sodas populaires utilisés dans cette étude, une seconde variable a été créée, codifiée comme l'inverse de la première variable respective, c'est-à-dire qu'un 1 a été saisi si la boisson respective n'a pas été consommée ni achetée, et 0 a été saisi si elle a été consommée ou achetée

au cours du mois. Ci-dessous, observez une liste partielle des données codifiées de cette manière, pour 8 sodas courants. Ouvrez le fichier de données Beverage.sta situé dans le répertoire Boissons.

1) Réalisez une ACM sur ces données et retrouver ainsi les principales conclusions données dans l'aide de Statistica :

*Il s'avère que toutes les boissons sont raisonnablement bien représentées par la solution à deux dimensions, mis à part Pepsi Light dont la valeur de Qualité est inférieure à 0,5.*
*Un examen attentif du graphique suggère que le premier axe oppose essentiellement les boissons allégées aux boissons classiques, alors que la seconde dimension oppose les colas aux autres sodas.*

*En outre, si vous examinez attentivement les statistiques des coordonnées de lignes, vous allez constater que les individus contribuant le plus à l'inertie de la seconde dimension sont les observations numéro 13 et 28. Ces points "définissent" presque à eux seuls la direction de la seconde dimension.*

2) Reprenez ensuite l'étude à l'aide du fichier Beverag2.sta, dont les données sont saisies sous forme de tableau protocole.

### 4.6.1 Exercice à rendre

Référence : Comme les données "produits bio" précédemment étudiées, les données présentées ici sont (à des retouches mineures près) celles qui sont accessibles sur le site personnel de Gilles Hunault, à l'adresse :http://www.info.univ-angers.fr/pub/gh/Datasets/datasets.htm

Pour comparer l'effet de deux médicaments antalgiques (d'où le nom ANTAL pour le dossier) et de leur association, on soumet à 40 patients ces traitements "en double aveugle", c'est-à-dire que ni le patient ni le docteur ne savent de quel traitement il s'agit. Les quatre traitements sont notés A, B, AB et P. AB est en fait l'association de A et B, P est un placebo (absence de médicament). On note dans échelle de temps repérée par les identificateurs de lignes la réaction des patients. Ainsi A1/4 signifie après 1/4 d'heure pour le traitement A, P2h signifie après 2 heures pour le placebo, etc. Les colonnes signifient respectivement douleur nulle (DOU0), douleur légère (DOU1), douleur modérée (DOU2), sévère (DOU3), très sévère (DOU4) ; on indique aussi une douleur théorique (DOU5) qui devrait être associée voire équivalente à DOU4.

Les données observées, présentées comme ci-dessus ont été saisies dans la feuille Antal du classeur Excel Antalgiques.xls. Elles sont également présentées sous une autre forme dans la feuille Antal1 du même classeur.

A l'aide d'AFC ou d'ACM, procédez à une analyse exploratoire des données observées. Quelles différences semble-t-il apparaître entre les 4 traitements du point de vue des variables étudiées (intensité de la douleur, et temps) ?

Vous pouvez utiliser les données sous l'une ou l'autre forme. Mais vous veillerez à traiter de façon pertinente la variable (ou modalité, selon la forme des données) DOU5, qui joue ici un rôle particulier.

Travail à rendre par mail à votre enseignant (carpenti@infolettres.univ-brest.fr si vous travaillez dans les salles de TD, Francois.Carpentier@univ-brest.fr si vous travaillez à l'extérieur) :

- Un classeur Statistica contenant les résultats numériques des analyses et les graphiques.
- Un fichier Word contenant une (brève) interprétation des résultats.