Licence d'Informatique 2
Analyse de Données Utilisateur (C5–160412)
TD 3 – Analyse de Données Multidimensionnelles

Carl FRÉLICOT – Dpt Info / Lab MIA

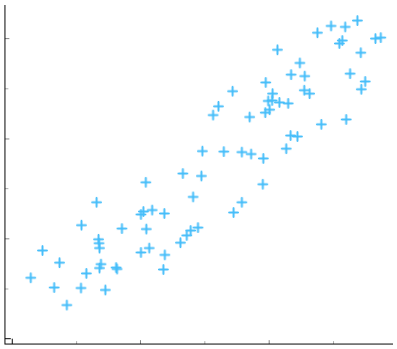
L'objectif de l'analyse multidimensionnelle est d'étudier simultanément un nombre p quelconque de variables :

- toutes quantitatives, ou — *Analyse en Composantes Principales* — [ACP]
- toutes qualitatives — *Analyse (Factorielle) des Correspondances (Multiples)* — [AFC, ACM]

Les méthodes sont essentiellement descriptives et consistent à rechercher des *facteurs* (ou *composantes principales*) en nombre restreint ($q < p$) qui résument le mieux le tableau de données. Ces facteurs sont des variables virtuelles[1] qui combinent linéairement les variables initiales, non corrélées deux à deux. On peut les utiliser pour visualiser le nuage de points initial dans l'espace de dimension inférieure engendré par les facteurs, plus pertinent.

1. Principe de l'ACP

Considérons un tableau de données bidimensionnelles ($p = 2$) représenté par son nuage de points ci-contre.



(c) Quelle est la ($q = 1$) direction principale du nuage ?
(d) Quelle statistique peut permettre de la qualifier ?
(e) Comment calculer les coordonnées de tous les points le long de cet axe ?
(f) Quelle pourrait être la 2ème direction principale ?

Imaginons un nuage en $p = 3$ dimensions.

(g) Essayez d'étendre le raisonnement précédent.

Exercice 1

Le tableau ci-contre représente les notes données par $n = 6$ critiques de cinéma à $p = 3$ films. Les moyennes et les écarts-type des notes des films sont donnés :

	Film 1	Film 2	Film 3
$\overline{x}$	6	4	5
s^2	10.66	5.67	5.67

	critique	Film 1	Film 2	Film 3
1	Cahiers	8.0	1.0	0.0
2	Libération	4.0	6.0	5.0
3	Cercle	6.0	8.0	7.0
4	TvHebdo	10.0	4.0	7.0
5	Télérama	8.0	2.0	5.0
6	Première	0.0	3.0	6.0

(a) Quelle(s) information(s) pertinente(s) pouvez-vous en extraire ?
(b) Calculez la somme des variances.

Le tableau ci-contre représente les $p = 3$ composantes principales obtenues. Les moyennes et variances des facteurs ont également été calculées :

	PC1	PC2	PC3
$\overline{x}$	-0.00	-0.00	0.00
s^2	12	8	2

	critique	PC1	PC2	PC3
1	Cahiers	4.899	3.464	1.414
2	Libération	-2.449	0.000	1.414
3	Cercle	-2.449	-3.464	1.414
4	TvHebdo	2.449	-3.464	-1.414
5	Télérama	2.449	-0.000	-1.414
6	Première	-4.899	3.464	-1.414

(c) Pourquoi les moyennes sont-elles nulles ?
(d) Calculez la somme des variances. Que remarquez-vous.
(e) Calculez le pourcentage cumulé des variances.
(f) Selon ce critère, quel nombre q de composantes peut-on retenir très bien représenter ces données ?

Les vecteurs directeurs des composantes principales sont donnés dans le tableau ci-contre :

	component	Film 1	Film 2	Film 3
1	PC1	0.816	-0.408	-0.408
2	PC2	-0.577	-0.577	-0.577
3	PC3	-0.000	0.707	-0.707

(g) Projetez un critique le long d'un vecteur directeur par produit scalaire.
(HW) Généralisez (pour tous les critiques) et toutes les composantes principales par produit matriciel. $C' = X'\,U$

où X' est le tableau de données centrées et U celui des vecteurs directeurs des composantes

[1] pour information, la recherche des facteurs est un problème de minimisation (d'un critère particulier) sous contraintes dont la résolution résulte en un ensemble de p *vecteurs propres*, orthogonaux deux à deux, et directeurs des axes qui définissent les facteurs. La théorie mathématique sous-jacente n'est pas absolument indispensable à la compréhension des méthodes d'analyse.

2. Analyse des Variables et des Individus

Trouver les facteurs est une chose, leur donner sens en est une autre. Pour cela, on peut bien sûr tenter de les expliquer à l'aide des variables initiales qu'ils combinent. De même, les individus participent plus ou moins à la formation des composantes principales.

(a) Comment quantifier si une composante principale est linéairement liée avec telle ou telle variable initiale ?
(b) Comment juger quel(s) individus sont bien représentés par une composante ?
(c) Selon vous, comment pourrait-on (numériquement) définir la *contribution* d'un individu à un facteur ?

Exercice 2

Les corrélations (linéaires) entre les composantes principales et les notes des films sont données dans le tableau ci-contre.

	PC1	PC2	PC3
Film 1	0.866	-0.500	-0.000
Film 2	-0.594	-0.686	0.420
Film 3	-0.594	-0.686	-0.420

(HW) Chez vous, vous retrouvez au moins une de ces valeurs.
(a) Tracez un cercle (dit des corrélations) de rayon 1, dont vous libellerez l'abscisse "PC1" et l'ordonnée "PC2".
(c) Dessinez le nuage de points-variables (les films) dans ce plan (PC1,PC2). Vous y porterez les *labels* des films.
(d) Sur une autre figure, dessinez le nuage de points-individus (PC1,PC2). Vous y porterez les sources des critiques.
(e) En vous appuyant sur l'analyse des variables précédente, sur les coordonnées des critiques sur le 1er facteur, et éventuellement sur le tableau initial de données, trouvez quelle information principale est portée par le 1er axe.
(f) Faites de même avec le 2ème facteur.
(g) Quelle information pourrait finalement porter le 3ème facteur ?

3. ACP normée

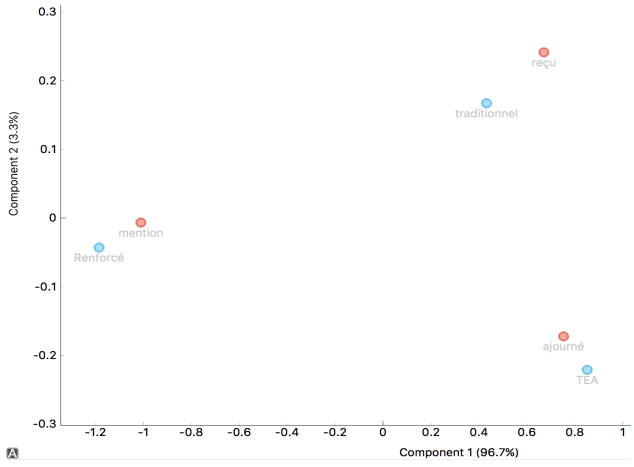
Une ACP est dite *normée* si au lieu d'être réalisée sur le tableau de données centrées, celles-ci sont centrées-réduites.

(a) Imaginons que parmi les p variables observées, une d'entre elles ait une variance très supérieure aux autres. Que va-t-il se passer lors d'une ACP non normée ?
(b) N'est-il alors pas plus prudent de systématiquement normer ?
(c) Sur quoi l'utilisateur doit-il s'appuyer pour décider s'il faut réaliser une ACP normée ou non normée ?
(d) Que vaut la somme des variances des variables centrées-réduites ? Celle des facteurs d'une ACP normée ?
(e) Déduisez un critère simple de choix du nombre q de composantes à retenir dans le cas d'une ACP normée.

4. Principe de l'AFC

Lorsque le tableau de données ne comporte que deux variables qualitatives X et Y, la méthode s'appelle *Analyse (Factorielle) des Correspondances* (AFC) ; s'il y en a plus, on parle d'*Analyse des Correspondances Multiples* (ACM).
À partir de la table de contingence croisant les modalités des deux variables, on peut définir le tableau des profils-ligne et celui des profils-colonnes (voir TD2). Le premier peut être vu comme un tableau de données croisant des individus (les modalités de X) et des variables (les modalités de Y), donc un nuage de points sur lequel on peut réaliser une ACP. De même le second tableau définit un nuage de points (les modalités de Y) décrit par des variables (les modalités de X). L'AFC consiste à réaliser cette double ACP est profils-ligne et des profils-colonne. On montre que ces deux ACP se correspondent, ce qui est normal puisque partant de la même table de contingence, et il est tout à fait légitime de superposer les axes ou les plans factoriels.

Il existe des indicateurs numériques permettant d'analyser à quel point les profils contribuent à la formation des axes, si ils sont bien représentés le long de ces axes, si les profils-ligne et les profils-colonne sont proches[a] ou non, et par conséquent s'il y a une *correspondance* (ou dépendance) entre des modalités de X et des modalités de Y. Ces indicateurs reposent sur des notions mathématiques qui peuvent dépasser un utilisateur...
Néanmoins, il peut tout à fait interpréter les sorties graphiques (voir TP).



[a] notez que les distances entre profils se ne sont pas des distances usuelles

5. Principe de l'ACM

L'*Analyse des Correspondances Multiples* (ACM) est l'extension de l'AFC à plus que deux variables qualitatives, ce qui change assez fondamentalement le tableau de données. Du point de vue d'un utilisateur, la façon d'interpréter les résultats d'un ACM sera cependant analogue à la façon d'interpréter ceux d'une AFC.