L'analyse en composantes principales en pratique

Après avoir vu sa formalisation mathématique dans le module précédent, on s'intéresse ici à l'utilisation pratique de l'ACP.

1 Objectifs

L'objectif de l'analyse en composantes principales (ou ACP) est purement descriptif : il s'agit « d'explorer » un ensemble d'observations rassemblées sous la forme d'un tableau de données indiquant pour chaque unité statistique les valeurs observées d'un certain nombre de variables quantitatives.

L'unité statistique peu être primaire (par exemple un individu, une entreprise) ou secondaire. Dans ce dernier cas, il s'agit de regroupement d'unités statistiques primaires.

Exemple d'unité statistique secondaire. On a interrrogé des clients d'un groupe bancaire sur la satisfaction de leur agence (notation de l'accueil, des heures d'ouverture, de la disponibilité, …). L'unité statistique primaire est l'individu. Si on veut travailler sur les agences du groupe bancaire (unité statistique secondaire) on calculera, par exemple, des moyennes de notation par agence pour chaque variable et on réalisera l'ACP en prenant comme individu les unités statistiques secondaires et comme variables les notes moyennes de chaque agence.

Le tableau de données peut être de dimensions importantes : le nombre de lignes (d'unités statistiques) peut atteindre plusieurs centaines, et le nombre de colonnes (de variables) plusieurs dizaines. Le nombre d'observations, suivant son importance, pourra donner un caractère de généralité aux propriétés structurelles ; il est en effet rare que l'on fasse appel, dans le cadre de l'analyse de données multidimensionnelle, à la statistique inférentielle.

L'analyse en composantes principales est fondée sur le calcul des moyennes, variances et coefficients de corrélation. Les données doivent donc être **quantitatives** : elles peuvent être discrètes ou ordinales (par ordre de préférence).

Exemple : On étudie les données sur 50 clients de l'hypermarché EUROMARKET constituées de l'âge, du revenu, du montant des achats, du nombre d'enfants, de la catégorie socioprofessionnelle (CSP) et du sexe. Les variables quantitatives sont les suivantes : l'âge, le revenu, le montant des achats, le nombre d'enfants. Nous verrons ultérieurement comment tenir compte du sexe et de la catégorie socioprofessionnelle dans les analyses.

Nous pouvons étudier les couples d'observations (âge, revenu) en les représentant graphiquement et en calculant le coefficient de corrélation. Cette représentation graphique montrera que le revenu s'accroît en fonction de l'âge, jusqu'à 60 ans environ, ce que nous pouvons expliquer par le fait qu'au-delà de 60 ans, les clients sont en retraite et voient leurs ressources financières diminuer.

L'analyse en composantes principales généralise cette démarche en prenant en compte la totalité des variables quantitatives : ainsi, nous verrons que les personnes de 60 ans et plus n'ont en général pas d'enfant à charge, et par suite le montant de leurs achats est moins élevé : il y a donc une tendance générale dans les données, liée à l'âge, qui permet d'expliquer la diminution de la consommation de plusieurs façons.

La taille de ce tableau est insuffisante pour que les interprétations soient intéressantes. Mais elle permet de donner la totalité des résultats concernant les variables et d'effectuer des calculs sur quelques unités statistiques à l'aide d'une simple calculatrice.

La taille du tableau de données rend les calculs à la main impossible et donc l'utilisation d'un logiciel de traitement spécifique est indispensable. Ces logiciels sont très nombreux et l'on peut les segmenter selon plusieurs types :

- les logiciels de traitement d'enquête (Le Sphinx, ethnos, Question, …). Bien que leur spécialité soit le traitement de questionnaires, ils intègrent quelques méthodes d'analyses

factorielles. Les sorties sont relativement sommaires et les options disponibles sont limitées (pas de rotation des axes, …)

- les logiciels boîtes à outils (XLSTAT, Statbox). Ils permettent de réaliser diverses analyses factorielles (ACP, AFC, ACM), quelques techniques de classification (Classification hiérarchique, K moyennes) ainsi que les techniques de prévision classiques. Les données sont gérées à partir du logiciel Microsoft Excel et les sorties s'effectuent dans des feuilles de calculs. Globalement, ils offrent un bon rapport qualité/prix
- Les logiciels de statistique (SPSS, SPAD, SAS, …). Conçus pour manipuler et analyser de grands tableaux de données, ils sont très complets sur le plan des méthodes présentes et sur les options disponibles. L'utilisation est plus complexe et nécessite parfois plusieurs journées (voire plusieurs mois) de formation. Leur prix en fait un outil réservé aux cabinets statistiques ou aux directions statistiques de grandes entreprises.

Dans ce cours, nous présenterons une sortie du logiciel STAT MANIA, et utiliserons pour les exercices et activités les sorties du logiciel Statbox.

2) La réalisation de l'ACP

Pour réaliser une ACP on suit une démarche en plusieurs étapes :

1 Préparation des données

S'assurer que les données sont **quantitatives**. Dans la pratique, on considère souvent les variables qualitatives ordinales comme des quantitatives. Par exemple, dans les enquêtes de satisfaction les variables qualitatives ordinales possèdent les modalités suivantes :

Pas du tout satisfait ; plutôt pas satisfait ; moyennement satisfait ; plutôt satisfait ; Tout à fait satisfait.

On considère que ces modalités correspondent a une note donnée par l'individu avec 1 pour Pas du tout satisfait, 2 pour plutôt pas satisfait … et on obtient ainsi que variable quantitative discrète que l'on pourra utiliser en ACP.

Remarque : en pratique on s'autorise une certaine liberté d'interprétation qui n'a pas de fondement statistique. En effet, 4 est supérieur à 2, ce qui traduit bien que « plutôt satisfait » indique une satisfaction supérieure à « plutôt pas satisfait ». Mais, mathématiquement, 4 est le double de 2 ; et rien ne justifie le fait que « plutôt satisfait » traduise une satisfaction deux fois plus importante que « plutôt pas satisfait ».

On rappelle également que la variable sexe, même si elle est codifié 1 pour les hommes et 2 pour les femmes est une variable qualitative et ne doit donc pas être utilisée dans l'ACP.

Données manquantes : L'ACP ne sait pas traiter les données manquantes. Certains logiciels proposent de supprimer les individus possédant des données manquantes, alors que d'autres vont remplacer la donnée manquante par un zéro.

2 Paramétrer le logiciel

Il faut indiquer au logiciel les divers paramètres de l'ACP :

- les variables actives (celles qui permettront de discriminer les individus),

- les variables supplémentaires (voir § suivant),

- la présence éventuelle d'individus supplémentaires

- le nombre de valeurs propres a calculer

- le nombre d'axes à représenter

- éventuellement, le libellé des individus (ou l'identifiant des individus)

Individus et variables supplémentaires

Individus supplémentaires

Afin de faciliter l'interprétation des résultats, on peut introduire dans le tableau de données de départ des données que l'on appelle individus supplémentaires.

Les unités statistiques supplémentaires sont des unités statistiques sur lesquelles on dispose des observations des variables mais dont on ne veut pas tenir compte dans le calcul des paramètres statistiques. On définit souvent comme unités statistiques supplémentaires les centres de gravité de groupes formés à priori, définis par les moyennes des variables de ces groupes. Ainsi, dans l'exemple d'Euromarket, on pourrait introduire dans les données précédentes deux individus supplémentaires l'un caractéristique du groupe « Homme » et l'autre du groupe « Femme »

Le tableau de données de départ devient donc

N°	âge	revenu	achats	enfants
1	51	195888	150.15	3
2	39	128456	173.12	2
…	…	…		
GHommes	Age moyen des hommes	Revenu moyen des hommes	Achat moyen des hommes	Nb d'enfants moyen des hommes
GFemmes	Age moyen des femmes	Revenu moyen des femmes	Achat moyen des femmes	Nb d'enfants moyen des femmes

On définit donc deux unités statistique supplémentaires dont on ne doit pas tenir compte dans les calculs puisqu'ils ne représentent pas d'unités statistiques réelles : ce sont des unités statistiques supplémentaires.

L'intérêt des données supplémentaires est de caractériser sur les graphiques des groupes d'unités statistiques supplémentaires

Variables supplémentaires :

Ce sont des variables n'ayant pas de rapport direct avec l'analyse mais que l'on souhaite voir représentées dans les graphiques.

Certains logiciels utilisent les termes de variables actives et passives (pour supplémentaires).

3 Réaliser les calculs

On a vu dans les modules précédents les différents calculs à réaliser. Vu la taille du tableau de données que l'on traite habituellement, c'est le logiciel qui réalisera cette étape.

Le logiciel produit alors différents tableaux et graphiques (mapping) qu'il faudra interpréter.

3) Interpréter les résultats

1 Déterminer le nombre d'axes de l'analyse

Pour répondre à cette question, il faut consulter le tableau des valeurs propres qui accompagne l'ACP. Les valeurs propres sont classées de façon décroissante. L'inertie de chaque axe et l'inertie cumulée figurent également dans ce tableau.

Il y a deux manières pour déterminer le nombre d'axes à prendre en compte :

- Un critère "absolu" : ne retenir que les axes dont les valeurs propres sont supérieures à 1 (c'est le critère de Kaiser).

- Un critère "relatif" : retenir les valeurs propres qui "dominent" les autres, en se référant au graphique en barres des valeurs propres ("screeplot", chez les Anglo-saxons).

Il est important que les valeurs propres des axes retenus restituent une "bonne proportion" de l'analyse. Cela signifie que la somme de l'inertie expliquée par chacun des axes représente une partie importante de l'inertie totale. Cette somme est une mesure de la fiabilité de la lecture des mappings, et donc de la qualité globale explicative de l'analyse.

3 Sélectionner les individus et variables à interpréter

Les mapping de l'ACP sont les projections des variables et des individus sur un plan factoriel déterminé. On commencera par interpréter le premier plan factoriel (celui formé par les facteurs F1 et F2) car c'est celui qui concentre la plus grande partie de l'information du nuage. On ira voir ensuite et le cas échéant les autres plans factoriels.

Sur un plan factoriel, on n'interprète que les variables et les individus qui sont bien représentés. Pour les individus, on utilisera les contributions absolues et relatives alors que pour les variables, on n'interprètera que celles qui sont proches du cercle de corrélation.

4 les sorties graphiques

Deux mapping sont données par les logiciels : celui des variables et celui des individus

La représentation des variables.

Ce mapping se distingue par la présence d'un cercle de corrélation. Sur un plan factoriel déterminé, on n'interprète que les variables qui sont bien représentées c'est à dire celles qui sont proches ou sur le cercle de corrélation. On interprète deux types de positions :

Les positions des variables par rapport aux axes afin de déterminer quelles sont les variables qui « font les axes ». On va ainsi pouvoir nommer les axes en fonction des variables.

Les positions des variables les unes par rapport aux autres. Le coefficient de corrélation entre deux variables étant le cosinus de l'angle formé par les vecteurs on en déduit que :

- deux variables qui sont proches ou confondu (angle de 0°) sont corrélées positivement (coefficient de corrélation proche de 1),
- deux variables opposées (formant un angle de 180°) sont corrélées négativement (coefficient de corrélation proche de -1)
- deux variables positionnées à angle droit (angle de 90°) ne sont pas du tout corrélées (coefficient de corrélation égal à 0)

La représentation des individus

Deux cas se présentent :

L'ACP est réalisé sur un tableau comportant beaucoup d'individus (plus de 30). Dans ce cas, on ne pourra pas interpréter les positions relatives de tous les individus car le nuage sera tellement dense que l'on n'y verra pas grand-chose. Toutefois, si un individu est atypique, il va ressortir du nuage et on pourra alors l'identifier pour éventuellement le supprimer et effectuer un nouveau passage sans cet individu. Dans ce cas, on a souvent recours à une méthode classification automatique afin de regrouper les individus qui sont proches les uns des autres et ainsi de constituer des type d'individus ayant un comportement similaire.

Si l'ACP est réalisée sur un nombre d'individus plus faible, l'interprétation du nuage des individus est alors possible. C'est notamment le cas lorsque l'on travaille avec des unités statistiques secondaires, où il sera par exemple intéressant d'étudier la position de telle ou telle agence bancaire.

Quel que soit le cas envisagé, on n'interprète, sur un plan factoriel déterminé, que les individus qui sont bien représentées. Pour cela, il faut aller voir leurs contributions absolues et relatives.

Sous réserve d'une bonne représentation, la proximité de deux individus sur un plan factoriel est synonyme d'individus ayant un comportement similaire, c'est-à-dire ayant des réponses aux variables de l'analyse qui sont très proches. Si deux individus ont exactement les mêmes valeurs aux

différentes variables, ils seront superposés sur les différents plans factoriels. De même, des individus ou des groupes d'individus s'opposant par rapport à un axe factoriel, s'opposeront par rapport aux variables qui « font » cet axe.

> *Remarque : variables et individus supplémentaires apparaissent généralement sur les mapping avec une couleur différentes des variables et individus actifs.*

La représentation individus/variables

Bien qu'elle soit mathématiquement contestable, les logiciels standard fournissent une représentation graphique dans laquelle est juxtaposée la projection des variables et des individus. Cette représentation permet de visualiser les individus ayant des valeurs élevés (ou faible) de telle ou telle variable. (voir l'exemple suivant)

Un exemple de sortie du logiciel STATMANIA

1 Le Tableau de données

Le tableau de données est constitué des notations moyennes des clients pour les différents magasins du groupe (Paris, Lyon, Marseille, Nice ….)

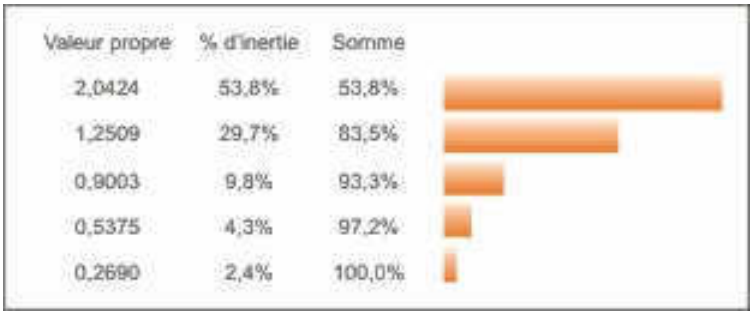
Les variables sont : Choix proposés, Facilité pour trouver le produit, Disponibilité des vendeurs, Compétence des vendeurs, Courtoisie des vendeurs

L'unité statistique est donc le magasin. (en ligne dans le tableau)

Les sorties de l'ACP sont celles du logiciel Statmania .

2 Nombre d'axes de l'analyse

Exemple de tableau de valeurs propres du logiciel STAT MANIA



Valeur propre	% d'inertie	Somme
2,0424	53,8%	53,8%
1,2509	29,7%	83,5%
0,9003	9,8%	93,3%
0,5375	4,3%	97,2%
0,2690	2,4%	100,0%

Sur le schéma précédent on remarque qu'en conservant les deux premiers axes on va expliquer 83,5% de l'inertie totale du nuage de point.

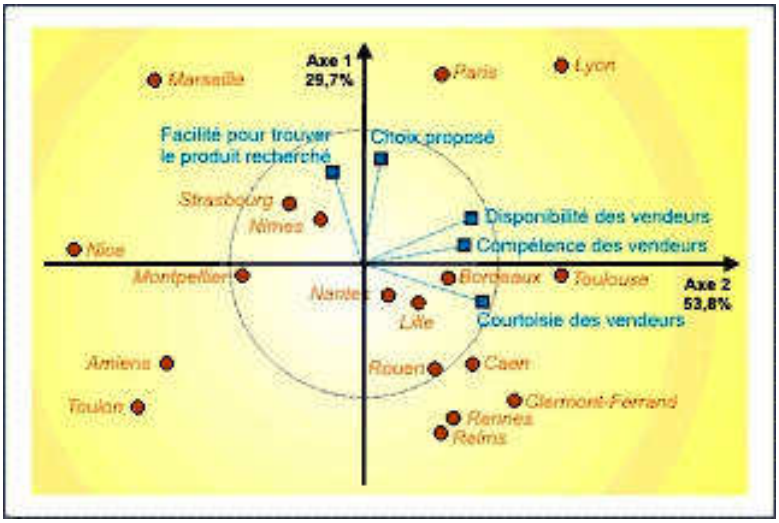
2. Quels sont les points qui nous intéressent ?

Les points les plus intéressants sont généralement ceux qui sont assez proches d'un des axes, et assez loin de l'origine. Ces points sont bien correlés avec cet axe et sont les points explicatifs pour l'axe : Ce sont les points les plus "parlants" ; leur "vraie distance" de l'origine est bien représentée sur le plan factoriel[1].

Dans le mapping ci-dessous, on voit clairement que Nice est extrêmement corrélé avec l'axe horizontal. De même, Paris et Reims notamment sont très bien correlés à l'axe vertical.

[1] Mais rien n'empêche d'aller vérifier que leur contribution soient bonne.

La corrélation de chaque point sur un axe exprime la qualité de représentation du point sur l'axe. Elle prend des valeurs entre 0 (pas corrélé du tout) et 1 (fortement corrélé). Si cette valeur est proche de 1, alors le point est bien représenté sur l'axe.



Les points situés près du centre sont donc généralement mal représentés par le plan factoriel. Leur interprétation ne peut donc pas être effectuée avec confiance. Ainsi Nîmes et Strasbourg semble proche mais c'est peut être le fruit d'une projection car ils sont peut être opposés sur l'axe 3.

Pour être plus rigoureux, il faut interpréter les points en fonction de leurs CTR et CTA. Ces valeurs sont généralement données par les logiciels.

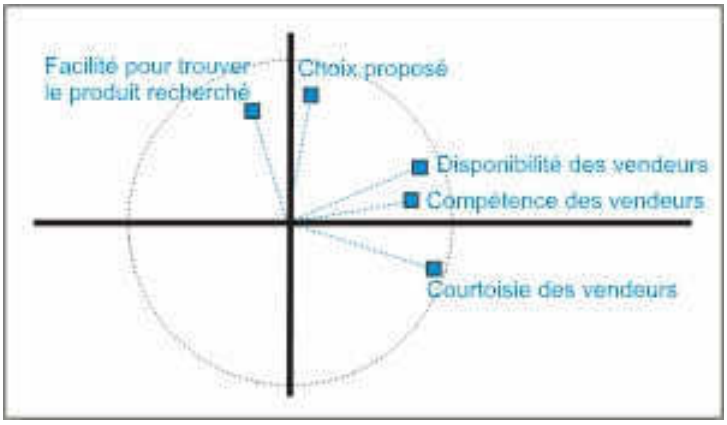
3. Comment interpréter les proximités ?

On s'intéresse donc essentiellement aux points bien représentés (i.e. situés loin du centre). Si deux points sont proches l'un de l'autre, il est probable que les réponses des individus qu'ils représentent soient très similaires. Il faut cependant se méfier :

Il se peut que sur un axe ils soient très proches, alors que sur un autre ils seront très loin l'un de l'autre. Il faut donc les regarder par rapport à tous les axes qui ont été retenus pour l'analyse. S'ils sont bien corrélés avec l'axe qui les montre proches, alors, on peut conclure qu'ils sont vraiment proches.

Est-ce qu'on peut donner un sens "réel" aux axes du mapping ?

Les axes factoriels sont des axes virtuels issus d'une synthèse entre les variables de l'analyse. Ils n'ont pas nécessairement un sens précis même si on peut souvent leur trouver un sens en s'aidant notamment de la représentation des variables sur le cercle de corrélation. Rappelons que la représentation de ce cercle et des variables sur le mapping de l'ACP se fait sur une échelle arbitraire, ce qui implique que la proximité des points variables par rapport aux points individus n'a absolument aucun sens.



Dans notre exemple, nous pouvons constater que les points "disponibilité", "compétence" et "courtoisie" sont très proches du cercle de corrélation et donc très bien représentés sur le mapping. L'angle plutôt fermé (en partant de l'origine) que forment les points "compétence" et "disponibilité" indique que ces 2 variables sont assez bien corrélées entre elles. En revanche, l'angle quasi droit formé par "compétence" et "choix" indique que ces deux variables sont indépendantes entre elles.

Le fait que "compétence" soit proche de l'axe 1 indique qu'il est très bien représenté par cet axe. Comme il est très éloigné de l'axe 2, on peut conclure qu'il est peu représenté par cet axe.

En ce qui concerne l'axe 2, le point "choix" est très bien correlé avec l'axe. Le point "facilité" l'est également mais dans une moindre mesure.

De ces observations, nous pouvons conclure que l'axe 1 correspond plutôt à l'appréciation des vendeurs et notamment de leur compétence alors que l'axe 2 correspond plutôt à l'appréciation du magasin et notamment du choix qu'il propose.

Quelles autres conclusions tirer de notre analyse ?

En synthétisant les informations issues des 5 variables analysées, notre mapping nous montre qu'il y a beaucoup d'efforts à faire en matière d'accueil et de renseignement des clients dans les magasins de Nice, Marseille, Amiens et Toulon. Ce dernier est également très peu apprécié en matière de choix.

Les magasins de Paris, de Lyon et de Marseille sont appréciés de la clientèle pour le choix qu'ils proposent et la facilité pour trouver les produits recherchés.

Lyon se distingue aussi par l'amabilité du personnel et peut être considéré comme le meilleur magasin parmi ceux qui ont fait l'objet de l'analyse.

Ces conclusions sont confirmées par l'examen des tableaux de corrélations et de coordonnées des individus, fournis par le logiciel d'analyse.