5 Réduction de dimension. Analyse factorielle. ACP

Pour ce chapitre on vous demande d'utiliser le fichier de données point_de_vente_12_12_2023_ACP

Table des matières

1 Objectifs....1
2 La mesure de l'information. 2
3 La procédure....2
4 Espace des individus et espace des variables....3
5 Extraction et rotation....3
6 Hypothèses sur les données....3
7 L'ACP avec Jamovi....4
8 Facteurs à retenir....4
9 Analyse des variables....4
9.1 Graphique des variables....4
9.2 Tableau des variables....5
9.3 Interprétation....5
10 Analyse du nuage des individus....5
10.1 Graphique des individus....5
10.2 Tableaux des individus....6
10.3 Interprétation....6
11 Lien entre l'analyse du nuage des individus et celle du nuage des variables....6
12 Utilisation des facteurs. 6

1 Objectifs

Un questionnaire porte sur un grand nombre d'aspects. Dans l'exemple du point de vente, il y a 23 variables. Il est difficile d'appréhender ces 23 dimensions. Il est courant d'avoir beaucoup plus de 23 variables.

L'analyse en composantes principales ou ACP est une méthode d'analyse factorielle, c'est-à-dire qu'elle se propose de synthétiser des variables en quelques-unes seulement. Elle s'applique à des variables quantitatives.

A partir de K **variables** initiales, dont chacune correspond à une question d'un questionnaire, elle crée de **nouvelles variables**, appelées **facteurs.** Chacun de ces facteurs est une combinaison linéaire des variables initiales, c'est-à-dire qu'il est obtenu en multipliant chacune des variables initiales par un coefficient fixe et en ajoutant le tout :

$$facteur = a_1\, variable_1 + a_2\, variable_2 + ... + a_K\, variable_K$$

Ces facteurs sont construits pour qu'un petit nombre d'entre eux suffise à porter une part suffisante de l'information portée par les variables initiales.
Ainsi l'analyse factorielle permet

- de résumer les données en **créant** des variables qui **n'apparaissaient pas** au départ mais qui

synthétisent l'information

- de **réduire la dimension** des données.

2 La mesure de l'information.

On considère des individus décrits par K variables numériques.
S'il n'y a que 2 variables quantitatives, on peut représenter graphiquement les données par un nuage de points dans un plan en reportant chaque variable sur un axe.

Pour 3 variables, on peut utiliser une représentation en 3 dimensions, utilisant 3 axes.

Pour k > 3 variables, il faut imaginer disposer de k axes … dans un espace à k dimensions ! Ce qui est difficile car nous sommes habitués à notre espace physique à 3 dimensions... Néanmoins cela se fait et on peut calculer des moyennes et des distances dans des espaces de plus de 3 dimensions.

En particulier on peut calculer un point moyen G, dont la valeur pour chaque variable n° i serait la moyenne des valeurs de tous les individus , $\overline{x_i}$. On peut ensuite évaluer pour chaque individu n°j , une distance entre G et cet individu ainsi : $d_j = \sqrt{\sum_{i=1}^{K} \left(x_i^{(j)} - \overline{x_i}\right)^2}$ où $x_i^{(j)}$ est la valeur de la variable n°i pour cet individu n°j

On considère que l'information portée par l'ensemble des points est mesurée par I = $\frac{1}{N}\sum_{j=1}^{N} d_j^2$: plus les points sont éloignés les uns des autres, mieux on les distingue .

Si on ne considère pour chaque individu que p variables, p < K – par exemple on ne s'occupe que des p premières questions et on oublie les autres –

- ➔ on représente les données dans un espace à p dimensions, plus facilement compréhensible.
- ➔ Mais on distingue moins bien les individus : on ne connaît plus leurs différences sur les (K-p) variables qu'on a exclues. Pour chacun d'eux on ne perçoit plus entre lui et G que la distance suivante :

  $d_j' = \sqrt{\sum_{i=1}^{p} \left(x_i^{(j)} - \overline{x_i}\right)^2} \leq d_j$ car $d_j = \sqrt{\sum_{i=1}^{p} \left(x_i^{(j)} - \overline{x_i}\right)^2 + \sum_{i=p+1}^{K} \left(x_i^{(j)} - \overline{x_i}\right)^2}$, donc l'information portée par le nuage des points diminue : $I' < I$
  .

3 La procédure

1. Si les données comprennent K variables, l'analyse factorielle crée K nouvelles variables, que l'on peut représenter sur K axes … d'un espace à K dimensions.
   Ces **nouvelles variables** sont appelées des **facteurs**. Pour les construire
   - on commence dans la grande majorité des cas par centrer et réduire les variables initiales, c'est-à-dire par leur soustraire leur moyenne et les diviser par la racine carrée de leur variance. Après cette transformation **les variables ont toutes pour moyenne 0 et pour variance 1**, et sont devenues des nombres sans dimension, ce qui est bien pratique pour comparer des variables de natures différentes – par exemple des sommes d'argent, des poids ou des longueurs ….– .
   - Chaque facteur est défini comme une **combinaison linéaire des variables initiales centrées et réduites**, c'est-à-dire que ces variables sont multipliées chacune par une constante, puis additionnées. Dans le cas du point de vente une combinaison linéaire des variables de départ est par exemple : 2 × déco + 3 × originales – 0,5 × taille_foyer.

- Ces combinaisons linéaires sont choisies de façon à ce que la première variable porte une **part maximale** de **l'information**, la seconde porte une information maximale tout en étant représentée par un axe orthogonal à la première – dans un espace à K dimensions ...-- et ainsi de suite, les nouvelles variables sont orthogonales dans l'espace à K dimensions et elles portent une information décroissante...

2. L'analyse factorielle applique ensuite un critère pour ne **retenir parmi les K facteurs, que les plus importants.**

Parmi trois critères couramment utilisés, en voici deux

- On ne retient que les facteurs qui portent au moins autant d'information que les anciennes en moyenne. Cela se traduit dans l'algorithme par "**on retient les variables ayant une valeur propre supérieure ou égale à 1**". Cet algorithme utilise des matrices, dont les valeurs propres traduisent les quantités d'information portées par les axes de représentation.
- On retient des facteurs qui, dans leur ensemble, portent une quantité suffisante d'information – c'est à dire **une part suffisante de la variance** (70%, 80 % …) – .

La dimension des données est ainsi diminuée, tout en conservant autant que possible la quantité d'information.

4 Espace des individus et espace des variables.

Le tableau de données point_de_vente_12_12_2023_ACP contient 68 lignes de 30 colonnes pour décrire 68 individus par 30 caractères chacun. On peut considérer que

- Chaque individu est décrit par une ligne, les 30 valeurs des variables pour cet individu, et donc peut être représennté dans un espace de 30 dimensions.
- Chaque variable est décrite par une colonne, contenant les 68 valeurs des 68 individus pour cette variable. Elle peut se représenter dans un espace à 68 dimensions. En fait, puisqu'elles ne sont que 30, elles sont contenues dans un sous-espace de 30 dimensions « seulement ».

L'ACP construit donc 30 facteurs pour représenter les individus dans leur espace et 30 facteurs pour représenter les variables dans l'espace des variables, qui n'est pas celui des individus. De plus ces facteurs sont fortement liés. Lorsqu'il sera question de valeurs propres un peu plus loin, la valeur propre attachée au facteur n°j de l'espace des variables est égale à celle du facteur n°j de l'espace des individus. Par la suite, «le kième facteur » pourra désigner celui de l'espace de variables ou celui de l'espace des individus.

5 Extraction et rotation

On dit que **les étapes précédentes ont extrait les facteurs** à retenir. On pourra ensuite créer des graphiques pour lesquels chaque axe de coordonnées porte les valeurs d'un des facteurs. Mais ces facteurs, combinaisons linéaires de variables diverses, n'ont pas forcément d'interprétation facile. L'analyse factorielle procède ensuite à une seconde opération : on **applique des rotations** aux axes portant les premiers facteurs pour définir un **second ensemble de facteurs**, de façon à ce que chaque variable initiale ne soit fortement corrélée qu'avec un seul d'entre eux. Ces facteurs issus de la rotation sont plus facilement interprétables : ils mesurent les variables qui leur sont fortement corrélées.

6 Hypothèses sur les données

L'ACP est une méthode descriptive, et non probabiliste. Elle ne demande **aucune condition probabiliste sur les données.** En particulier les variables n'ont pas besoin de suivre une loi de Gauss, ni d'être indépendantes – au contraire on recherche des variables corrélées, donc non indépendantes. –.
Cela étant, **on peut** émettre des hypothèses probabilistes pour construire des tests statistiques .

7 L'ACP avec Jamovi.

On utilise le module MEDA qui fournit plus d'indications que le menu Facteurs de Jamovi. S'il n'est pas installé, vous pouvez le faire en cliquant sur la croix « modules » en haut à droite de la fenêtre Jamovi.

Si ce n'est pas fait, ouvrir le fichier « Point_de_vente_12_12_2023_ACP »

Cliquer sur MEDA, puis « Principal component analysis »

Introduire dans « Active variables » les variables : conseil, originales, choix, classique, rock, naissance, rap, marques, sobre.

Dans « Graphic options » et « Numerical indicators » vous pouvez cocher toutes les cases

8 Facteurs à retenir

Le tableau « Eigenvalue decomposition » indique les valeurs propres – traduction de eigenvalue – des différents facteurs. Ce sont aussi les inerties, ou encore les variances, portées par ces facteurs. Les facteurs sont numérotés de 1 à n, nombre de variables initiales, et désignés par les noms Dim 1 à Dim n.

Il est aussi indiqué quel pourcentage de la variance – ou de l'inertie – totale est portée par chaque facteur,

et le pourcentage cumulé de variance – ou d'inertie – porté par les premiers facteurs.
On peut donc **juger du nombre de facteurs à considérer**, en utilisant le critère de la valeur propre – donc de l'inertie, et de la variance – supérieure à 1 , ou celui du pourcentage cumulé de valeur propre – ou d'inertie, ou de variance – supérieure ou égale à 80%.

9 Analyse des variables

On dispose de graphiques et d'un tableau.

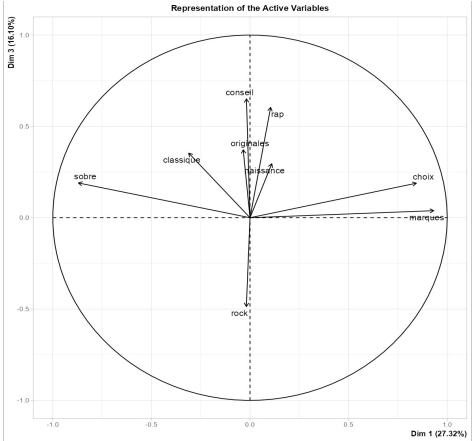
9.1 Graphique des variables

Voici le graphique des variables dans un repère où l'axe des abscisses représente les valeurs du 1er facteur – celui ayant la valeur propre la plus élevée et portant le plus d'information – et l'axes des ordonnées celles du 2nd facteur. Les 2 premiers facteurs portant le maximum d'information sur l'ensemble des variables, c'est celui à consulter en premier. Mais on va voir qu'on peut juger sur quels facteurs une variable précise est représentée au mieux, et on peut tracer un graphique avec ces facteurs pour axes, pour cela

1. dans la rubrique « Numerical indicators --> automatic description of the axes --> number of factors » indiquer le nombre de facteurs retenus.
2. dans la rubrique « Graphic Options –> Components to plot » indiquer les numéros des facteurs à représenter dans les cases « X-axis » et « Y-axis ».

Sur ce graphique :

- Deux variables **corrélées** seront représentées sur le graphique par deux vecteurs de directions proches, de même sens pour une corrélation positive, de sens opposé pour des variables corrélées négativement.
- Attention, il n'y a pas de réciproque : il s'agit de la projection en 2 dimensions de vecteurs d'un espace de pusieurs dimensions. De la même façon qu'une photographie en 2 dimensions d'une grenouille en gros plan devant vache, vivant toutes les deux dans notre espace en 3 dimension, pourrait faire penser que la grenouille a une taille comparable à la vache, il est possible que des vecteurs de directions proches sur ce graphique représentent des variables éloignées en réalité.



- Deux variables décorrélées sont représentées par des vecteurs orthogonaux.
- Ce sont bien les directions des vecteurs et non leurs extrémités qu'il faut comparer. Les variables ont été normées au début de la procédure, le fait que leurs représentations planes aient des longueurs différentes dépend du plan selon lequel on les regarde.
- Les variables ont toutes pour norme 1, leur projections apparaissent toujours à l'intérieur d'un cercle unité.
- Plus l'extrémité du vecteur représentant une variable dans un plan est proche du cercle de rayon 1, et mieux cette variable est représentée dans ce plan.

Question 1 Quelles sont les variables les mieux représentées sur ce plan ?

Question 2 Il **semble** se dégager 4 groupes de variables, quelles sont-elles ?

9.2 Tableau des variables.

Coordinates table donne les coordonnées des variables dans une base des facteurs. La coordonnée d'une variable sur un facteur est égale au coefficient de corrélation de la variable et de ce facteur. Les variables et les facteurs étant de norme 1, et les facteurs étant orthogonaux, la coordonnée d'une variable sur un facteur est aussi égale au cosinus de l'angle entre cette variable et ce facteur, dans l'espace des variables. Et bien sûr la variable est égale à la somme des facteurs multipliés chacun par la coordonnée de la variable sur ce facteur.

Contribution table indique dans quelle mesure une variable a contribué à la construction d'un facteur dans l'algorithme de l'ACP. Plus il est grand et plus la variable est influente dans la détermination de la direction de cet axe.

Cosine est le carré du cosinus de l'angle entre la variable et le facteur. Plus il est proche de 1 et mieux la variable est représentée par ce vecteur.

Pour juger de la bonne représentation d'une variable *v* dans un plan dont les axes sont des facteurs F_1 *et* F_2, on utilisera $\left(\cos(F_1, v)\right)^2 + \left(\cos(F_2, v)\right)^2$

9.3 Interprétation.

On peut dégager des **groupes de variables** ainsi :

On repère pour chaque facteur n°k retenu dans l'ACP les variables fortement corrélées avec cet axe. La corrélation étant égale à la coordonnée sur un facteur, on peut se fonder sur le graphique des variables, ou sur le tableau des coordonnées. On retient dans un groupe K+ les variables dont la corrélation avec ce facteur est supérieure à 0,70 et dans un autre, K-, celles dont la corrélation est inférieure – 0,70.

Les variables du groupe K+ sont corrélées positivement entre elles et ont tendance à prendre de grandes valeurs en même temps que le facteur K. Elles s'opposent à celles du groupe K- , qui ont tendance à prendre de grandes valeurs lorsque celles du facteur K sont faibles. Il faut se fonder sur la signification des variables pour donner un sens à ces groupes et aux facteurs.

10 Analyse du nuage des individus.

10.1 Graphique des individus

- Sur un nuage de points des individus dans un repère où figurent en abscisse et en ordonnée des facteurs d'une ACP, des individus proches ont des représentations proches – à la différence des variables où ce sont les directions des représentations qui sont proches.

- Attention, ici aussi il n'y a pas de réciproque : du fait de la dimension des données supérieure à 2, des représentations proches n'indiquent pas forcément des individus proches. Avant d'en juger on peut vérifier s'ils sont bien représentés par les facteurs qui sont en abscisses et en ordonnées; le tableau *cosine* des individus fournit le carré du cosinus de l'angle entre chaque facteur et la représentation de l'individu. Un individu est bien représenté dans un plan factoriel si la somme des 2 cosinus carrés est supérieure ou égal à 0,70.

10.2 Tableaux des individus.

Cosine fournit le carré du cosinus de l'angle entre le facteur et le vecteur $\overrightarrow{GM}$, G point moyen du groupe. M point réprésentant l'individu.

Coordinates fournit la coordonnée de chaque représentation d'individu sur chaque facteur.

Contributions Table fournit le pourcentage de contribution du point à la construction de l'axe.

10.3 Interprétation.

Comme pour les variables, on peut créer pour chaque facteur n°k deux groupes ainsi : on regroupe les individus de plus grande contribution à cet axe, et dont la somme des contributions atteint 80% . On-distingue le groupe k+ des individus qui parmi eux, ont une coordonnée positive sur le facteur et le groupe k- de ceux qui ont une coordonnée négative sur l'axe.

Les individus du groupe k+ prennent plutôt de grandes valeurs sur ce facteur et s'opposent que groupe k- qui prennent plutôt de petites valeurs.

11 Lien entre l'analyse du nuage des individus et celle du nuage des variables.

Comme cela a déjà été indiqué, les facteurs de l'espace des individus et ceux de l'espace des variables ont un lien. Et il se trouve que

pour chaque facteur numéro k,

Les individus du groupe k+ ont de grandes valeurs des variables du groupe de variables **de même numéro** K+ et de faibles valeurs des variables du groupe K- . Ils s'opposent aux individus du groupe k- dont c'est l'inverse.

12 Utilisation des facteurs.

Les facteurs sont des variables que vous pouvez mémoriser dans votre fichier de données en cochant « save coordinates » dans « Numerical indicators » de la fenêtre de commande . Elles apparaissent avec les noms DIM1, DIM2, … Vous pouvez les utiliser comme n'importe quelles autres variables de vos données.

Bibliographie : L'analyse multivariée avec SPSS, Stafford , Bodson, 2006 ;

Analyse factorielle multiple avec R, Pagès, 2013. Disponibles en ligne sur le site de la bibliothèque.

Site du logiciel libre TANAGRA : https://tanagra-machine-learning.blogspot.com/

Site de l'université de Sherbrooke consacré à l'utilisation de SPSS : https://spss.espaceweb.usherbrooke.ca/