Analyse factorielle en composantes principales

Michel Calciu

Cours à l'Université de Lille 1 - 2011/2012

Introduction

Réduction et positionnement

L'analyse en composantes principales a pour objet la description synthétique de tableaux de données dans lesquels des individus sont décrits par des variables quantitatives multiples. Cette description doit permettre :

- une réduction de l'information; les variables descriptives sont regroupées au sein de facteurs synthétiques, les composantes principales, qui correspondent à des dimensions sous-jacentes du problème;

- le positionnement des individus par rapport à ces composantes principales, ce qui peut mettre en évidence des typologies d'individus ainsi que les variables qui ont amené à la création de ces types[1].

L'étude d'un échantillon ou d'une population ne peut prétendre habituellement être complète que si un nombre élevé de variables, appelées critères, tests, ou mesures sont évaluées pour chacun des cas. L'ensemble de ces mesures couvre de façon complète, du moins on l'espère, une partie structurée et connaissable du domaine d'investigation. À première vue, chacune de ces variables pourrait sembler d'égale importance; considérant cependant que plusieurs d'entre elles sont en corrélation, donc redondantes, il est possible de découvrir l'existence d'un plus petit nombre de variables dans un ordre décroissant d'importance, indépendantes (du moins habituellement) les unes des autres et telles que les premières expliquent la plus grande partie de la dispersion. C'est l'objectif que se proposent les divers modèles de l'analyse factorielle et qu'ils atteignent, du moins dans leurs grandes lignes[2].

Modèle géométrique à deux variables

Données brute

Voici une situation à deux variables permettant d'illustrer géométriquement le but poursuivi par la recherche des composantes principales. Supposons un ensemble de sujets mesurés sur deux variables.

Prenons le cas des differentes modèles de voitures.

Obs	Modèle	Prix	Cylindrée
AS2	Austin Métro Special	39999	998

CI4	Citroën AX 10 RE	44250	954
DA2	Daihatsu Charade 1000 TS	48750	993
FI3	Fiat Panda 1000 Cl.	40333	999
FI5	Fiat Uno 45 fire	44916	999
FI8	Fiat Uno Turbo TE	83350	1301
FID	Fiat Uno 70 SL	59483	1302
FO1	Ford Fiesta Junior	43500	1117
FO9	Ford Fiesta X R-2	72476	1597
NI1	Nissan Micra 1.0 DX	41333	988
OP1	Opel Corsa Swing Belux 1.0 S	43500	993
PE1	Peugeot 205 XE 1.0	44200	954
PE3	Peugeot 205 GL	52600	1124
PE6	Peugeot 205 GT	63216	1360
PE9	Peugeot 205 GTI	83916	1580
RE1	Renault 4 TL	43270	956
RE3	Renault 4 GTL	46020	1108
RE4	Renault 5 SL	47187	1108
RE7	Renault 5 GTS	57500	1397
RE8	Renault 5 GT Turbo	84395	1397
SE4	Seat Ibiza 1.5 GLX	64278	1461
SE9	Seat Marbella 900 GL	36565	903
SZ2	Suzuki Swift 1.0 GA	40383	993
SZ3	Suzuki Swift 1.3 GL	48850	1324

TO1	Toyota Starlett 1000 L	46666	999
TO2	Toyota Starlett 1300 XL	56166	1295
VW3	Volkswagen Polo Coupé GT	60150	1272

Figure

Dans un espace à deux dimensions, chaque sujet est représenté par un point dont les coordonnées sont les scores (figure 1).

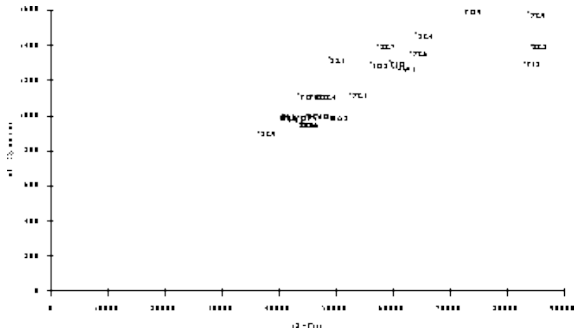


Figure 1. Contour elliptique de deux variables conjointes normalement distribuées.

Données centrées

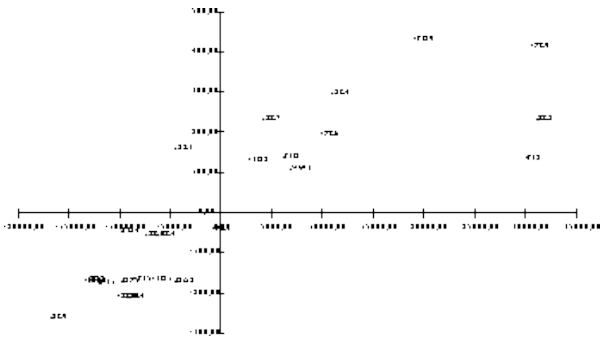
Pour centrer les données on soustrait de chaque élément d'une colonne du tableau (chaque valeur d'une variable) la moyenne de la colonne (de la variable).

Obs	Modèle	Prix	Cylindrée
AS2	Austin Métro Special	-13232,56	-167,63
CI4	Citroën AX 10 RE	-8981,56	-211,63
DA2	Daihatsu Charade 1000 TS	-4481,56	-172,63
FI3	Fiat Panda 1000 Cl.	-12898,56	-166,63
FI5	Fiat Uno 45 fire	-8315,56	-166,63
FI8	Fiat Uno Turbo TE	30118,44	135,37

FID	Fiat Uno 70 SL	6251,44	136,37
FO1	Ford Fiesta Junior	-9731,56	-48,63
FO9	Ford Fiesta X R-2	19244,44	431,37
NI1	Nissan Micra 1.0 DX	-11898,56	-177,63
OP1	Opel Corsa Swing Belux 1.0 S	-9731,56	-172,63
PE1	Peugeot 205 XE 1.0	-9031,56	-211,63
PE3	Peugeot 205 GL	-631,56	-41,63
PE6	Peugeot 205 GT	9984,44	194,37
PE9	Peugeot 205 GTI	30684,44	414,37
RE1	Renault 4 TL	-9961,56	-209,63
RE3	Renault 4 GTL	-7211,56	-57,63
RE4	Renault 5 SL	-6044,56	-57,63
RE7	Renault 5 GTS	4268,44	231,37
RE8	Renault 5 GT Turbo	31163,44	231,37
SE4	Seat Ibiza 1.5 GLX	11046,44	295,37
SE9	Seat Marbella 900 GL	-16666,56	-262,63
SZ2	Suzuki Swift 1.0 GA	-12848,56	-172,63
SZ3	Suzuki Swift 1.3 GL	-4381,56	158,37
TO1	Toyota Starlett 1000 L	-6565,56	-166,63
TO2	Toyota Starlett 1300 XL	2934,44	129,37
VW3	Volkswagen Polo Coupé GT	6918,44	106,37

Figure

Le graphique pour les premières deux variables est illustré en figure 2



Données centrées et réduites

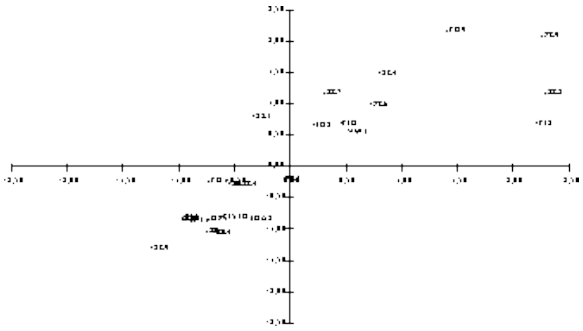
Souvent pour éviter les grandes différences d'ordre de grandeur entre les variables on réduit les donne en divisant chaque colonne du tableau (chaque valeur d'une variable) par l'écart type de la colonne (de la variable).

Obs	Modèle	Prix	Cylindrée
AS2	Austin Métro Special	-0,98	-0,84
CI4	Citroën AX 10 RE	-0,66	-1,06
DA2	Daihatsu Charade 1000 TS	-0,33	-0,86
FI3	Fiat Panda 1000 Cl.	-0,95	-0,83
FI5	Fiat Uno 45 fire	-0,61	-0,83
FI8	Fiat Uno Turbo TE	2,22	0,68
FID	Fiat Uno 70 SL	0,46	0,68
FO1	Ford Fiesta Junior	-0,72	-0,24
FO9	Ford Fiesta X R-2	1,42	2,15
NI1	Nissan Micra 1.0 DX	-0,88	-0,89
OP1	Opel Corsa Swing Belux1.0 S	-0,72	-0,86

PE1	Peugeot 205 XE 1.0	-0,67	-1,06
PE3	Peugeot 205 GL	-0,05	-0,21
PE6	Peugeot 205 GT	0,74	0,97
PE9	Peugeot 205 GTI	2,26	2,07
RE1	Renault 4 TL	-0,74	-1,05
RE3	Renault 4 GTL	-0,53	-0,29
RE4	Renault 5 SL	-0,45	-0,29
RE7	Renault 5 GTS	0,31	1,15
RE8	Renault 5 GT Turbo	2,30	1,15
SE4	Seat Ibiza 1.5 GLX	0,82	1,47
SE9	Seat Marbella 900 GL	-1,23	-1,31
SZ2	Suzuki Swift 1.0 GA	-0,95	-0,86
SZ3	Suzuki Swift 1.3 GL	-0,32	0,79
TO1	Toyota Starlett 1000 L	-0,48	-0,83
TO2	Toyota Starlett 1300 XL	0,22	0,65
VW3	Volkswagen Polo Coupé GT	0,51	0,53

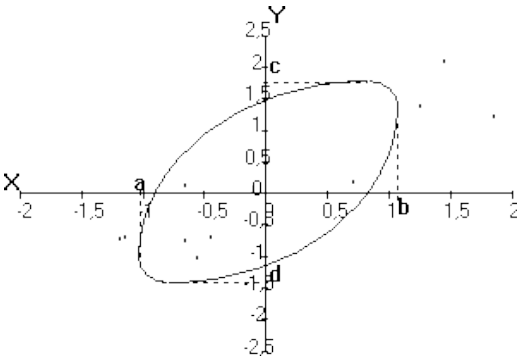
Figure

Le graphique pour les premières deux variables est illustré en figure 3:



Forme générale du nuage de points

La forme générale du nuage de points est celle d'une ellipse si les variables sont distribuées normalement. Une ligne contour d'égale densité permet de mieux voir la distribution de ces points.



Considérons les dispersions des deux variables X (Cylindrée) et de Y (Prix) qui sont mesurées habituellement par les variances, mais qu'on représente ici comme les projections extrêmes des lignes contours. Sur l'axe des X la dispersion va de a à b et sur l'axe des Y, elle va de c à d. On constate que les dispersions sur ces deux axes sont habituellement assez importantes pour qu'on doive tenir compte des deux variables pour expliquer adéquatement la dispersion totale: une explication simplifiée par réduction du nombre de variables ne semble donc pas, à première vue du moins, pouvoir être envisagée dans ces conditions. De plus, on observe une corrélation entre les deux variables: il y a redondance, c'est-à-dire que l'une des variables contient une partie de l'information de l'autre.

Recherche d’axes qui maximisent la variance

Considérons maintenant les deux axes de l'ellipse comme nouveaux axes de référence du système. Dans le cas de variables standard, ces nouveaux axes forment un angle de 45° par rapport aux anciens. Cet angle est différent dans le cas de variables non standard.

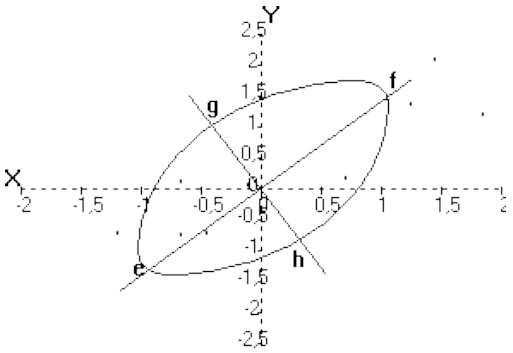


Figure 2. Représentation des sujets dans l'espace des axes de l'ellipse.

En effectuant une modification des coordonnées des points du nuage afin d'exprimer leurs positions par rapport aux nouveaux axes, on réalise une transformation intéressante. Ces nouveaux axes représentent de nouvelles variables, appelées composantes, fonctions des anciennes variables X et Y et dont la dispersion dans le nuage de points est maximum pour l'une, allant de e à f, et minimum pour l'autre, allant de g à h. Ainsi, la première composante rend compte de la plus grande partie de la variance alors que la deuxième composante, en plus d'être indépendante de la première, joue un rôle d'autant moins important que la corrélation $\mathbf{r}_{xy}$ est élevée; le rôle de la deuxième composante peut même dans certains cas devenir négligeable. L'indépendance des composantes ressort de l'orientation de l'ellipse relativement aux nouveaux axes de référence.

Le but qu'on se donnait est atteint: une situation décrite antérieurement par deux variables liées l'est maintenant par deux composantes indépendantes dont la première est plus importante qu'aucune des variables et dont la seconde, dans certains cas, peut être négligée.

L'importance de variables et composantes

L'*importance d'une variable ou d'une composante est donnée par sa variance* ou sa somme des écarts à la moyenne. On comprendra mieux intuitivement cet énoncé si on considère qu'à la limite, une variable dont tous les scores sont égaux, c'est-à-dire dont la variance est nulle, est une variable qui n'ajoute aucun renseignement à une situation.

La plus grande partie du chapitre portera sur l'analyse de variables standard.

C'est le cas le plus universel et celui où toutes les variables sont au départ considérées comme d'égale importance.

Plusieurs auteurs proposent l'extraction des composantes de variables centrées mais non réduites. Cette approche a comme point de départ la matrice des variances et covariances **V** plutôt que celle des corrélations **R**; ils vont même jusqu'à suggérer comme point de départ la matrice des sommes des carrés des écarts aux moyennes et des sommes des produits croisés. Cette méthode établit ainsi une hiérarchie des variables; cette hiérarchie est basée sur l'étendue des échelles des variables. Or l'étendue d'une échelle est arbitraire. D'autre part, les techniques d'analyse factorielle ont pour but l'explication des corrélations entre ces variables; et on sait que l'indice r de corrélation est totalement indépendant de l'étendue des échelles des variables. C'est pourquoi une approche autre que celle reposant sur la matrice de corrélations **R** qui impose de traiter toutes les variables sur un même pied, apparaît à plusieurs points de vue gratuite et injustifiée; le lecteur pourra s'en tenir à cette partie de l'exposé. On a cru bon cependant d'ajouter un article sur les conséquences de l'usage de variables non standard, c'est-à-dire sur l'extraction des

composantes à partir de la matrice V des variances et covariances, ce qui permettra d'évaluer les différences et similitudes des méthodes.

Modèle géométrique à plus de deux variables

Trois variables

Les sujets d'un échantillon mesurés sur trois variables liées peuvent être représentés dans un espace à trois dimensions par un nuage de points de forme ellipsoïdale; les lieux d'égale densité constituent des surfaces contours dont toute intersection avec un plan décrit une ellipse; un ballon de football plus ou moins aplati donne une bonne idée d'un tel contour. Les trois axes principaux sont perpendiculaires et représentent les trois composantes ordonnées. Ces trois composantes sont indépendantes; de plus la première rend compte de la plus grande partie de la variance du système tandis que la variance expliquée par les deux autres composantes est moindre. Dans certains cas, l'importance des deuxième et troisième composantes peut être faible au point de rendre le système presque totalement explicable par la seule première composante, ou dans les situations moins simples, par les première et deuxième composantes.

Plusieurs variables

On peut facilement imaginer la situation où un plus grand nombre de variables est étudié, ce qui entraîne espace multidimensionnel, nuage de points hyper-ellipsoïdal, hyperespace, hyperplan, etc.

La méthode des composantes principales est aussi dite des axes principaux, ce qui rend plus explicite le modèle géométrique. Un axe principal dans une ellipse est un segment qui, passant par le centre de l'ellipse, atteint celle-ci perpendiculairement à une tangente. Pour espace à trois dimensions et plus, les mots ellipse et tangente sont remplacés par ellipsoïde et plan tangent.

La recherche des composantes principales consiste à déterminer ce qu'on pourrait considérer comme les **longueurs** (racines latentes symbolisées $_j$) et les **directions** (vecteur fj des vecteurs latents) des axes principaux.

La première composante correspond à la meilleure description possible de la position des points du nuage. Cette exigence peut s'exprimer mathématiquement en fixant que la somme des carrés des écarts de chaque point à cet axe soit minimum, ou encore, en imposant que la variance de cette composante soit maximum. Chaque composante additionnelle répond au même critère et exige en plus d'être perpendiculaire à toutes les précédentes. L'orthogonalité observée des composantes répond à l'exigence de leur indépendance.

Développement du modèle algébrique

Introduction

On a vu que le modèle géométrique de la recherche d'une composante consiste à transformer les coordonnées de points, ou ce qui est équivalent, à effectuer une rotation orthogonale des axes, de telle sorte qu'on en arrive à une variance maximum du premier facteur. Le vecteur de transformation (ensemble des coefficients de saturations d'un facteur) agit sur la matrice des variables de façon à créer une variable

nouvelle appelée composante. On considérera dans cet article le cas où une importance égale est attachée à chaque variable, en transformant la matrice M des scores bruts en la matrice X des scores standard.

Position du problème

Le vecteur des scores de la composante répond à l'expression **y** = **Xf** où **f** est le vecteur de transformation recherché.

La variance de la composante s'écrit:

y'y/n = (**Xf**)'(**Xf**)/n =**f'X'Xf**/n = **f'Rf**

Cette variance est maximum pour **f** répondant à la condition:

$$\frac{\partial \boldsymbol{f}' R f}{\partial \boldsymbol{f}'} = 0$$

$$2\mathbf{Rf} = 0$$

$$\mathbf{Rf} = 0$$

Cas simplifié à deux variables

Examinons le cas simple, mais généralisable, d'une matrice X à deux variables; la condition du maximum de la variance sera:

$$\boldsymbol{Rf} = \{$$

c'est-à-dire: $a_1 + ra_2 = u$

$ra_1 + a_2 = 0$

d'où $(1 - r^2)\, a_1 = 0$ et $(1 - r^2)\, a_2 = 0$

On constate alors qu'une solution sans intérêt ($a_1 = a_2 = 0$) est la conséquence de

$(1 - r^2) = 0$ c'est-à-dire de r 1

On constate aussi qu'une solution indéterminée est la conséquence de la condition

$(1 - r^2) = = 0$ c'est-à-dire de r = 1

C'est une condition rarement rencontrée.

Utilisation du terme de Lagrange

Pour contourner cette difficulté, on ajoute à l'expression de la variance la condition **f'f=k** fixant ainsi les valeurs de **f** et du terme de Lagrange; ce qui donne la fonction:

V = **f'Rf-**(**f'f** - **k**).

La variance de la composante est maximum pour

= 2**Rf** - 2**f** = 0

d'où l'équation:

(**R** - **I**)**f** = 0

Revenons au cas particulier des deux variables; I'expression

(**R** - **I**)**f** = 0

peut s'écrire

= 0

d'où on tire

$[(1-)^2 - r^2]\ a_1 = 0$ et $[(1-)^2 - r^2]\ a_2 = 0$

La seule solution acceptable est celle de l'indétermination, c'est-à-dire de

$[(1-)^2 - r^2] = |\mathbf{R} - \mathbf{I}| = 0$

Cette condition détermine, d'une façon générale, le nombre de racines latentes possible: ce nombre est égal à l'ordre de **R**.

Resultat

A chaque valeur $_i$ correspond un vecteur $\mathbf{f}_i$obtenu par la solution de $(\mathbf{R} - {}_i\mathbf{I})\mathbf{f}_i = 0$. L'ensemble des $\mathbf{f}_i$ constitue la matrice **F** de transformation. La levée de l'indétermination se fera en fixant ou normant la valeur de **f**'j**f**j .

L'ensemble de ces conditions peut s'écrire:

RF = **F**

Application au marché des voitures

Presentation

Le tableau complet sur les modèles de voitures à sept variables[3], ça veut dire que les modèles analysées sont jugés selon sept critères. Pour reduire la complexité d'une telle comparaison on va utiliser l'analyse en composantes principales pour trouver un nombre reduit (deux) de composantes principales (indépendantes) qui captent la plus grande partie de l'information et qui remplaceront les sept variables, qui souvent sont corrélées entre elles.

Matrice de corrélations

On calcule la matrice de correlations (**R = X'X**/n) en multipliant la trasposée du tableau des données centrées et reduites avec le tableau. La matrice **R** calculée est la suivante:

1,00	0,85	-0,77	0,32	0,22	0,81	0,91
0,85	1,00	-0,78	0,30	0,11	0,80	0,83
-0,77	-0,78	1,00	-0,10	0,10	-0,68	-0,94
0,32	0,30	-0,10	1,00	-0,06	0,21	0,16
0,22	0,11	0,10	-0,06	1,00	0,29	0,02
0,81	0,80	-0,68	0,21	0,29	1,00	0,78
0,91	0,83	-0,94	0,16	0,02	0,78	1,00

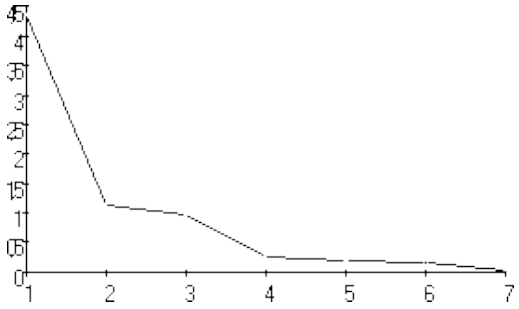
Sur cette matrice (**R**) on calcule les valeures et les vecteurs propres selon un algorythme connu :

Extraction des valeurs et vecteurs propres

4,355	1,123	0,97	0,232	0,172	0,134	0,014
0,4560	0,0740	0,0690	-0,2330	-0,0070	0,716	-0,4630
0,4440	-0,0170	0,0420	0,2670	0,8390	-0,133	0,0880
-0,4240	0,2260	0,2780	0,3500	0,1560	0,607	0,4210
0,1420	-0,2470	0,9270	-0,1260	-0,1180	-0,164	0,0500
0,0690	0,9090	0,1520	-0,3050	0,0310	-0,223	0,0450
0,4240	0,2120	0,0210	0,7420	-0,4660	-0,078	-0,0330
0,4590	-0,1040	-0,1830	-0,3040	-0,1980	0,134	0,7720

Utilisation des valeurs propres

La première ligne du tableau donne les valeures propres $_i$. La somme de valeures propres $_i$ = 7 et correspond à la variance totale (informations totale) des sept variables centrées reduites. Ca veut dire que en remplaceant les variables originales par les composantes principales on ne perd pas d'informations. En plus on constate que les premiers deux facteurs expriment 78,3% (5,5 / 7) de la variance (information totale).

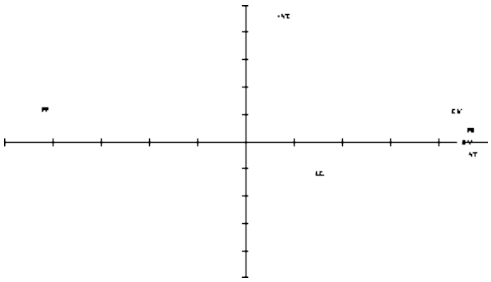


Ca veut dire qu'on peut sans trop de perte d'information utiliser seulement ces deux composantes comme axes pour representer le nuage de points et par rapport aux axes de variables utilisées dans les graphiques precedents ces axes sont ortogonales (noncorrélées entre elles).

Utilisation des vecteurs propres

En dessous des valeurs propres sont rangés en colonne les vecteurs propres qui sont les coefficients avec les quells sont pondérées les variables pour obtenir les facteurs. Pour voir l'orientation des variables par rapport aux composantes principales ils suffit de placer sur un grafique les coordonnées représentés par

les vecteurs propres:



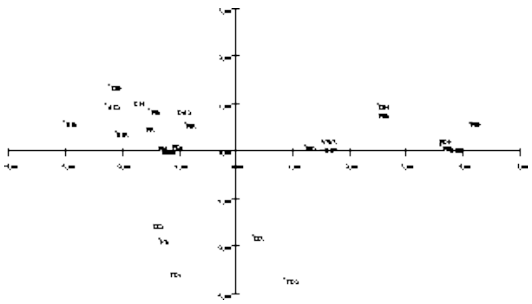
L'orientation des variables par rapport aux axes permet d'interpreter ces axes. Les variables qui sont plus proches du première axe du coté positif comme la vitesse, le prix , la cylindrée et du coté négatif le poids/puissance, permentent une iterpretations d'itermpreter l'axe comme étant l'axe de performances (techiques..), l'autre axes auquels sont corrélés positivement le volume du coffre et négativement la logueur.

Calcul de scores factoriels

Les nouvelles coordonnées des observations sur les axes sont calculées conformément à la définition des facteurs (y = xf) en multipliant le tableau de données originales (centrées reduite) par la matrice des vecteurs propres, Le nouveau tableau est:

$$\mathbf{Y = XF}$$

Les coordonnées des modèles de voitures qui ce trouvent dans les colonne du tableau Y qui corespondent au premières deux axes sont ilustré dans l'image suivante;



Ainsi on obtient la carte de positionnement des sujets analysés, qui permet d'identifier des groupes et d'essayer d'interpreter.

Propriètes des vecteurs latents

Introduction

On a vu en détail au chapitre ll que la résolution de l'équation $\mathbf{RF} = \mathbf{F}$ qu'on écrit aussi $(\mathbf{R} - {}_i\mathbf{I})\mathbf{f}_i = 0$, entraîne des valeurs indéterminées des éléments de $\mathbf{f}_i$. Cette indétermination est levée en normant ces vecteurs. Les normes à l'unité et à ${}_i$ sont particulièrement utiles. Examinons-en les propriétés et les conséquences sur la variance des composantes. On représentera par $\mathbf{F}_1$ la matrice des vecteurs latents normés à l'unité et par F celle des vecteurs latents normés à ${}_i$.

A) Normes des vecteurs latents

1) Rappelons la propriété suivante des vecteurs latents normés à l'unité (le théorème 2 de l'article 2.6.4, Laforge):

$\mathbf{F'}_1\mathbf{F}_1 = \mathbf{I}$ (1)

Considérant les relations

$\mathbf{RF}_1 = \mathbf{F}_1$

et $\mathbf{F'}_1\mathbf{F}_1 = \mathbf{I}$

on voit que $\mathbf{RF}_1 = \mathbf{F}_1\mathbf{I} = \mathbf{F}_1(\mathbf{F'}_1\mathbf{F}_1) = (\mathbf{F}_1\mathbf{F'}_1)\,\mathbf{F}_1$

et donc que $\mathbf{F'}_1\mathbf{F}_1 = \mathbf{I}$ (2)

Les relations (1) et (2) établissent que la somme des carrés des éléments d'une même colonne de F1 est égale à l'unité, de même que celle des éléments d'une même ligne. Pour la matrice

$\mathbf{F}_1 =$

on aura SC

SC

ainsi que $a_1b_1 + a_2b_2 = 0$ et $a_1a_2 + b_1b_2 = 0$

2) Dans le cas d'une norme à $_i$ on sait que

$\mathbf{F'F} =$

Considérant les relations

$\mathbf{RF} = \mathbf{F}$ et $\mathbf{F'F} =$

on voit que $\mathbf{RF} = \mathbf{F(F'F)} = \mathbf{(FF')F}$

d'où $\mathbf{F'F} = \mathbf{R}$

Dans le cas d'une matrice de saturations

$\mathbf{F} =$

ces relations signifient

$c_1^2 + d_1^2 = 1$

$c_2^2 + d_2^2 = 1$

$c_1c_2 + d_1d_2 = r_{12}$

$c1^2 + c_2^2 = {}_1$

$d_1^2 + d_2^2 = {}_2$

$c_1d_1 + c_2d_2 = 0$

B) Variance des composantes

1) L'application de la matrice de transformation $\mathbf{F}_1$ à celle des variables standard $\mathbf{X}$ donne la matrice des composantes $\mathbf{Y}_1$:

$\mathbf{Y}_1 = \mathbf{XF}_1$

Les variances de telles composantes sont égales aux racines latentes ${}_j$. Rappelons que les variances des composantes sont

$\mathbf{Y'}_1\mathbf{Y}_1/n = \mathbf{F'}_1\mathbf{RF}_1 = \mathbf{F}_1\mathbf{F}_1 =$ (3)

2) L'application de la matrice de transformation $\mathbf{F}$ à celle des variables standard $\mathbf{X}$ donne la matrice des composantes $\mathbf{Y}$:

$\mathbf{Y} = \mathbf{XF}$

À la suite du théorème 4 de l'article 2.6.4, on montre que les variances de ces facteurs sont égales aux carrés des racines latentes. On a ici, d'après les relations précédentes, l'expression des variances des composantes:

$\mathbf{Y'Y}/n = \mathbf{F'RF} = \mathbf{F'F} = {}^{2}$

Pour rendre les variances égales à ce qu'elles sont dans le cas de la transformation unitaire $\mathbf{F}_1$ il est nécessaire de les diviser par . On écrit donc

$\mathbf{Y'}_1\mathbf{Y}_1/n = (\mathbf{F'RF})^{-1} = ({}^{-1/2}\mathbf{F'})\mathbf{R}(\mathbf{F}^{-1/2})$ (4)

Rapprochant les expressions (3) et (4), on observe donc la relation intéressante suivante entre $\mathbf{F}_1$ et $\mathbf{F}$:

$\mathbf{F}_1 = \mathbf{F}^{-1/2}$

On retient donc les relations symboliques $\mathbf{Y} = \mathbf{Y}_1 = \mathbf{Y}^{-1}$

On a pu constater, que $\mathbf{F}$ affiche dans l'immédiat plus d'informations intéressantes que $\mathbf{F}_1$ en particulier en ce qui a trait aux variances des composantes et aux corrélations entre les variables. C'est pourquoi on préfère souvent interpréter la nature des composantes à partir de $\mathbf{F}$ plutôt que de $\mathbf{F}_1$. Cependant, à l'occasion du calcul de scores factoriels qui respectent la somme initiale des variances, il est nécessaire d'utiliser la matrice de transformation unitaire $\mathbf{F}_1 = \mathbf{F}^{-1/2}$.

3) Les v variables de $\mathbf{X}$ sont standard alors que les composantes ont des variances égales aux racines latentes ${}_1, {}_2, \ldots$ On sait de plus que V = A, + A2 + -

$v = {}_1 + {}_2 + ..$

On pourra trouver utile dans certains cas de rendre standard ces composantes. Il suffit pour cela de diviser les éléments de chaque composante par l'écart type correspondant ${}_i^{1/2}$. D'où les composantes standard seront

$\mathbf{Y}_S = \mathbf{Y}^{-1/2} = \mathbf{XF}_1^{-1/2} = \mathbf{X}(\mathbf{F}^{-1/2})^{-1/2} = \mathbf{X}(\mathbf{F}^{-1}$.

C'est souvent cette dernière expression qu'on utilise dans les banques de programmes pour établir la

matrice des composantes standard ou des scores factoriels standard. Cette opération offre l'avantage de rendre comparables les divers scores d'un sujet et facilite l'identification des composantes.

Données sur le marché des petites voitures[4]

	Modèle	Prix	Cylindrée	Poids/Puissance	Longueure	Volume du coffre	Consommation	Vitèsse
AS2	Austin Métro Special	39999	998	23,2	3403	955	6,2	140
CI4	Citroën AX 10 RE	44250	954	19,4	3500	1170	5,6	145
DA2	Daihatsu Charade 1000 TS	48750	993	20,8	3610	1151	6,7	145
FI3	Fiat Panda 1000 Cl.	40333	999	21,8	3644	1088	6,3	140
FI5	Fiat Uno 45 fire	44916	999	21,5	3645	968	6,2	145
FI8	Fiat Uno Turbo TE	83350	1301	11	3644	968	8,9	200
FID	Fiat Uno 70 SL	59483	1302	16	3645	968	7,7	165
FO1	Ford Fiesta Junior	43500	1117	22,7	3645	900	7	137
FO9	Ford Fiesta X R-2	72476	1597	12	3645	973	9,3	180
NI1	Nissan Micra 1.0 DX	41333	988	17	3640	375	6,4	140
OP1	Opel Corsa Swing Belux 1.0 S	43500	993	22,4	3622	845	7,2	143
PE1	Peugeot 205 XE 1.0	44200	954	23,8	3705	1200	6,8	134
PE3	Peugeot 205 GL	52600	1124	21,4	3705	1200	5,8	142
PE6	Peugeot 205 GT	63216	1360	13,9	3705	1200	9,2	170
PE9	Peugeot 205 GTI	83916	1580	11,2	3705	1200	8,7	190
RE1	Renault 4 TL	43270	956	33,1	3670	950	6,3	115
RE3	Renault 4 GTL	46020	1108	28,4	3670	950	6,3	120
RE4	Renault 5 SL	47187	1108	20,6	3591	915	5,8	143

RE7	Renault 5 GTS	57500	1397	13,8	3591	915	7,9	167
RE8	Renault 5 GT Turbo	84395	1397	10,2	3591	915	8,7	200
SE4	Seat Ibiza 1.5 GLX	64278	1461	14,7	3637	1200	8,8	175
SE9	Seat Marbella 900 GL	36565	903	23,4	3475	1088	7,3	131
SZ2	Suzuki Swift 1.0 GA	40383	993	18,4	3585	400	6,4	145
SZ3	Suzuki Swift 1.3 GL	48850	1324	14	3585	400	6,5	163
TO1	Toyota Starlett 1000 L	46666	999	19,5	3700	202	6,1	150
TO2	Toyota Starlett 1300 XL	56166	1295	15	3700	202	6,8	170
VW3	Volkswagen Polo Coupé GT	60150	1272	14	3655	1040	8	170

Rotation des axes principaux

Inroduction

Il est rare qu'une analyse factorielle s'arrête à la détermination des composantes principales ou des facteurs communs. En effet, ces opérations donnent généralement lieu à une matrice des saturations où chaque variable est expliquée par plusieurs sinon toutes les composantes retenues. La définition des composantes est souvent difficile à cause du nombre plutôt grand de saturations élevées, en particulier dans le cas des premières composantes: on appelle composantes générales celles définies par toutes les variables et composantes de groupe celles qui ne le sont que par quelques-unes.

Soit un ensemble de quatre variables dont on a extrait les composantes. Les vecteurs de transformation, normés pour leurs racines latentes respectives, constituent la matrice F suivante:

Matrice des saturations

0,91	0,08	0,06	0,08	0,07	0,07	0,06
0,89	-0,02	0,04	0,07	0,07	0,06	0,06
-0,85	0,23	0,25	-0,07	-0,07	-0,06	-0,06
0,28	-0,25	0,84	0,02	0,02	0,02	0,02
0,14	0,93	0,13	0,01	0,01	0,01	0,01
0,85	0,22	0,02	0,07	0,07	0,06	0,06
0,92	-0,11	-0,16	0,08	0,07	0,07	0,06

1 2 3 4 5 6 7

4,36 1,12 0,97 0,36 0,34 0,31 0,29

Les sommes des carrés des éléments des lignes sont approximativement égales à 1: ce sont les variances de variables standard. Les sommes des carrés des éléments des colonnes sont égales aux racines latentes, variances des composantes.

On ne retient, pour l'illustration en deux dimensions, que les deux premières composantes symbolisées A et B: ce sont d'ailleurs les deux seules dont la variance est supérieure à l'unité, c'est-à-dire à celle des variables. Les proportions de la variance totale expliquées par ces deux facteurs sont respectivement

4,36/7 = 0,62 et 1,12/7 = 0,16; la matrice F" des saturations est ainsi réduite à la suivante:

Matrice F de saturations normé aux valeurs propres

	Composante	Composante
Variables	A	B
Prix	0,91	0,08
Cylindrée	0,89	-0,02
Poids/Puissance	-0,85	0,23
Longueure	0,28	-0,25
Volume du coffre	0,14	0,93
Consommation	0,85	0,22
Vitèsse	0,92	-0,11

Interpretation graphique

Dans la figure 1 on illustre les positions relatives des composantes et des variables. Les composantes étant indépendantes, on les représente par des axes orthogonaux A et B, tandis que les variables sont disposées comme des vecteurs dont les projections sur les axes A et B sont les saturations. chaque vecteur a pour longueur la racine carrée d'une somme de carrés de ligne de la matrice des saturations.

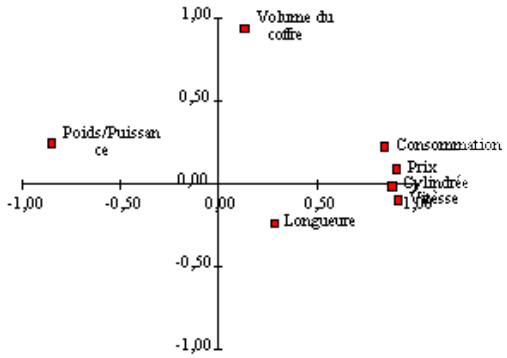


Figure 1. Longueurs et orientations des quatre variables et des facteurs A et B.

La nature de la composante A est définie par l'apport des sept variables à sa variance; ces **contributions** sont de 19,3% (0,91^2/4,36 = 0,193) pour la première variable, 18,17% ($0,89^2$/4,36 = 0,1817) pour la seconde variable, de 16,6% pour la troisième ($-0,85^2$/4,36 = 0,1657), de 1,8% pour la quatrième ($0,28^2$/4,36 = 0,018), de 4,5% pour la cinquième ($0,14^2$/4,36 = 0,0045) , de 16,6% pour la sixième ($0,85^2$/4,36 = 0,1657) et de 19,4% pour la septième ($0,92^2$/4,36 = 0,1941).

Ces contributions, lorsqu'elles sont importantes, ne permettent d ignorer aucune des variables: on dit d'une telle composante qu'elle est générale La détermination de sa nature serait cependant plus facile si les saturations fortes étaient moins nombreuses. On arrive à ce résultat en effectuant une rotation des axes de référence jusqu'en une position rendant maximum certaines saturations et minimum les autres.

Une rotation orthogonale

La figure 2 illustre une telle rotation orthogonale:

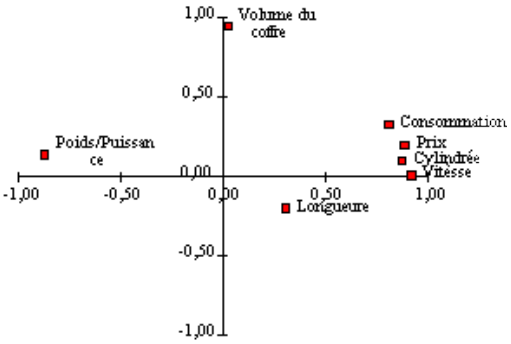


Figure 2. Projection des variables sur les nouveaux axes de référence .

Les nouvelles projections constituent la nouvelle matrice de saturations:

Variables	Composante C	Composante C
Prix	0,89	0,19
Cylindrée	0,88	0,09
Poids/Puissance	-0,87	0,13
Longueure	0,31	-0,21
Volume du coffre	0,02	0,94
Consommation	0,81	0,32
Vitèsse	0,92	0,00

Les deux composantes C et D ne sont pas les mêmes que les précédentes A et B, car elles constituent de nouvelles combinaisons linéaires de quatre variables. La composante C est définie surtout par les variables 1, 2, 3, 5 et 6, alors que D l'est par les variables 4 et éventuellement 5. Les nouvelles composantes sont alors dites de groupe par opposition aux composantes générales d'avant rotation.

Discussion

On a donc vu que la configuration de la structure factorielle n'est pas unique: une matrice de saturations définissant une structure peut être transformée sans en trahir les propriétés mathématiques et les hypothèses fondamentales. Il existe donc plusieurs moyens mathématiquement équivalents de définir les dimensions sous-jacentes à un même ensemble de données. Cependant ces solutions ne sont pas toutes équivalentes quant à leur degré de signification dans un domaine théorique donné; certaines respectent mieux la loi de la parcimonie scientifique, d'autres permettent une meilleure compréhension du domaine étudié. Le chercheur a donc la responsabilité finale du choix du type de rotation à effectuer.

En termes géométriques, les résultats recherchés dans une rotation peuvent se traduire ainsi: certaines des variables seront rapprochées de l'un ou de l'autre axe nouveau et auront sur ceux-ci des projections élevées; en même temps, elles feront avec les autres axes un angle voisin de 90° s'y projetant faiblement; le plus petit nombre possible de variables restera également éloigné des axes (voir Kim, dans SPSS. D. 484).

En d'autres termes, étant donné un nombre de facteurs expliquant une fraction fixe de la variance, il s'agit

de simplifier les lignes (méthode quartimax) ou les colonnes (méthode varimax) en rendant voisines de zéro le maximum de saturations.

Rotations varimax et quartimax

Il existe plusieurs procédés et critères mathématiques pour effectuer ces rotations. La méthode quartimax, proposée vers 1950 par plusieurs auteurs, consiste à maximiser la variance des carrés; comme cette méthode exige la maximisation de la somme des saturations à la quatrième puissance, on l'appelle quartimax. Une autre méthode proposée par Kaiser en 1958 repose sur la maximisation de la somme des variances des carrés des saturations dans chaque colonne il s'ensuit une augmentation de certaines saturations et la diminution des autres. Cette méthode dite varimax est la plus largement employée. Une troisième méthode, dite équimax, vise à la simplification simultanée des lignes et des colonnes de la matrice des saturations.

La rotation orthogonale de type varimax est obtenue au moyen de fonctions trigonométriques en traitant deux composantes à la fois:

=

La matrice de transformation est bien orthogonale car:

$\cos^2 + \sin^2 = 1$ et sin cos - sin cos = 0

Le traitement par ordinateur se prête à des rotations successives d'angle jusqu'à la satisfaction des critères suivants:

a) rotation varimax: $n(a_{ij}/h_i)^4$ - ^{2}

b) rotation quartimax: a_{ij}^4 maximum.

Comme on a pu le constater dans les figures précédentes, une rotation orthogonale des axes ne permet pas toujours de faire passer ceux-ci aux centres des agglomérations de variables. On pourrait y arriver cependant en effectuant plutôt une rotation dite oblique: on s'assure alors au maximum de l'augmentation ou de la diminution de certaines saturations: ce qui facilite l'interprétation des résultats. Cependant il est bien évident qu'on perd alors l'indépendance des facteurs. De plus, la démarche mathématique d'une telle méthode est plus complexe que celle d'une rotation orthogonale et certaines propriétés de la matrice des saturations sont perdues, dont celle de la somme des carrés des saturations de ligne qui n'est plus égale aux communautés. Après avoir connu une certaine faveur, les rotations obliques semblent laisser place chez les utilisateurs aux rotations orthogonales qui, plus simples à plusieurs points de vue, donnent des résultats satisfaisants. Nunnally (p. 321-333) offre une bonne discussion sur ces questions.

Exemple de rotation des axes principaux

Exemple de Thurstone

Le but de la rotation des axes est de faciliter l'identification de la nature des facteurs. Cependant la tâche n'est souvent pas facile et requiert une bonne connaissance de la nature des variables soumises à l'analyse

factorielle

On propose ici un exemple d'analyse emprunté à la géométrie élémentaire. Les résultats étant connus, comme le suggère Thurstone, on peut alors porter son attention sur la stratégie qui permet l'identification des facteurs.

L'objet de l'étude est le parallélépipède rectangle dont on veut découvrir les facteurs, c'est-à-dire les éléments géométriques essentiels à sa description. On a choisi de créer 100 boîtes dont les 3 dimensions, longueur (L), hauteur (H) et profondeur (P), devraient se dégager comme facteurs; ces dimensions sont générées de façon aléatoire. Pour ces 100 cas, on crée 1û variables y compris celles des dimensions, que l'on soumet à l'analyse factorielle par la méthode de Hotelling. La matrice de saturations, dont on ne retient que les trois premières colonnes, est la suivante pour les autres détails au sujet de

cette application):

Matrice de saturations (avant rotation)

		Facteur 1	Facteur 2	Facteur 3
Xl	longueur (L)	0,81	-0,16	-0,53
X2	hauteur (H)	0,61	-0,59	0,48
X3	profondeur (P)	0,51	0,76	0,36
X4	2(L+ P)	0,91	0,36	-0,15
X5	LH	0,84	-0,46	- 0, 1 2
X6	HP	0,73	0,02	0,66
X7	LP	0,86	0,31	-0,26
X8		0,88	0,40	-0,09
X9	LHP	0,92	-0,09	0,09
Xl0	2(L+H)	0,87	-0,44	-0,06

On constate rapidement que le nombre élevé de saturations importantes rend difficile l'identification des facteurs. Il peut même paraître étonnant que ces facteurs ne soient pas plus clairement mis en évidence par les trois premières variables: cela provient du fait que c'est en tenant compte des 10 variables que l'identification doit se faire.

-Cette matrice de saturations est alors soumise à une rotation orthogonale de type varimax, dont voici les résultats:

Matrice de saturations après rotation varimax

Matrice de saturations (rotation varimax)

		Facteur 1	Facteur 2	Facteur 3
Xl	longueur (L)	0,97	0,10	0,15-
X2	hauteur (H)	0,19	-0,01	0,96
X3	profondeur (P)	0,01	0,99	0,06
X4	2(L+ P)	0,68	0,70	0,15

X5 LH	0,75	0,01	0,61
X6 HP	0,05	0,60	0,78
X7 LP	0,74	0,60	0,08
X8	0,62	0,73	0,15
X9 LHP	0,60	0,41	0,58
Xl02(L+H)	0.73	0 05	0.65

Discussion

L'examen des trois premières lignes montre bien que les variables L, H et P sont les facteurs: chaque variable explique bien un et un seul facteur. Il est rare qu'en recherche on arrive à identifier à l'avance les facteurs; on se retrouve plutôt en présence d'une série de variables plus ou moins liées, comme le sont par exemple les variables composites 4 à 10 de notre tableau. L'identification de chaque facteur, à partir des saturations, est encore possible à la condition de bien connaître la composition des variables. En toutes circonstances, c'est de la connaissance approfondie des variables que dépend le succès de l'opération d'analyse factorielle; on ne saurait trop insister sur cette partie proprement créatrice de l'analyse où les ressources de connaissance et de réflexion du chercheur sont essentielles: le modèle mathématique et l'ordinateur ont fourni leur aide et ne sont plus d'aucun recours pour cette étape très spécifique au domaine étudié.

Le premier facteur est celui qui répond à l'ensemble des conditions suivantes: il est présent dans les variables 4, 5, 7, 8, 9 et 10 mais absent de la variable 6; c'est évidemment la longueur L. On identifie de façon semblable les deux autres facteurs. Le facteur 2 est présent dans les variables 4, 6, 7, 8, 9 mais absent des variables 5 et 10. L'examen de la composition, c'est-à-dire de la formule de ces variables, permet d'identifier comme étant P ce second facteur. De façon semblable, on identifie comme hauteur H le troisième facteur.

Remarques générales sur l'analyse factorielle

Conclusions

Il y a un certain nombre de précautions à prendre à l'occasion de l'emploi de l'analyse factorielle. En voici quelques-unes.

Quoiqu'en principe les scores factoriels soient indépendants, il n'en est pas toujours ainsi en pratique, spécialement si les communautés ou éléments de la diagonale de la matrice des corrélations sont différents de l'unité. Une autre raison de cette indépendance imparfaite est qu'un score factoriel est obtenu par la combinaison linéaire d'un nombre de variables inférieur à celui du problème original.

L'analyse factorielle a pour but d'étudier les corrélations "naturelles" entre plusieurs variables. Si ces corrélations étaient artificiellement obtenues, il ne faudrait pas s'étonner de voir apparaître une configuration artificielle de facteurs: ce serait le cas, par exemple, si une variable additionnelle était une combinaison linéaire d'autres variables, ou encore si les mêmes items apparaissaient dans l'élaboration de plus d'une variable; on rencontre une telle situation dans la construction des diverses échelles du MMPI (voir Shure et Miles, p. 14-18).

Il peut arriver que deux variables, dont les saturations sont élevées pour un facteur donné, soient effectivement sans corrélation entre elles. On peut observer une telle situation dans la matrice des saturations avant rotation dans l'article 8.4.2. Il est prudent, au moment de la définition des facteurs, de veiller à ne retenir que les variables qui ont entre elles des corrélations significatives.

Il est important de tenir compte de l'homogénéité des sujets. La structure factorielle peut être considérablement affectée par l'âge, le sexe, le niveau socio-économique, l'éducation des sujets. Si

l'homogénéité ne peut pas être facilement réalisée, certains auteurs préconisent l'insertion de l'une ou l'autre de ces influences comme variables additionnelles, à la condition de les annuler ensuite par l'emploi de la méthode d'analyse factorielle dite de la racine carrée (voir Nunnally, p. 370-371).

On ne retient en général que les composantes principales les plus importantes, c'est-à-dire expliquant la plus grande partie de la variance. Il peut arriver cependant que cette plus grande partie de la variance d'un sous-groupe soit attribuable aux composantes négligées: la discrimination entre ces sujets serait apparente surtout sur ces composantes Cette remarque suggère l'importance que peut revêtir la recherche des composantes principales pour divers sous-groupes de l'échantillon original.

Les composantes principales, étant indépendantes les unes des autres, sont plus faciles à interpréter que les variables elles-mêmes. Ce qui ne signifie nullement que cette interprétation est aisée. Habituellement on gagne beaucoup à joindre à cette analyse celles de la régression multiple et de la variance à variables multiples. Il ne faut point perdre de vue qu'une technique d'analyse factorielle n'est que l'application d'un modèle mathématique sur un ensemble de données numériques, dans le but de guider l'exploration préliminaire d'un domaine complexe.

On peut dans un rapport de recherche se contenter de ne présenter que les matrices des corrélations et des saturations. Une colonne de la matrice des saturations définit une composante. On ne doit pas cependant interpréter les composantes comme uniquement celles d'un certain nombre de tests; elles sont aussi les composantes de scores sur ces tests Lorsqu'on a défini une composante par l'examen d'une des colonnes de saturations, on peut procéder à sa vérification sur les scores factoriels de certains sujets Un sujet possédant la caractéristique de cette composante doit alors présenter un score factoriel proportionnel On peut procéder de façon plus systématique en regroupant les sujets, puis en les comparant au moyen de l'analyse de la variance

Si les mêmes facteurs ou composantes ont été obtenus sur plusieurs échantillons, on peut leur accorder une meilleure confiance L'analyse factorielle est habituellement précédée d'hypothèses, et c'est dans ce cas surtout que les facteurs sont utilisés comme concepts théoriques; sinon les composantes qui ne sont qu'une structure mathématiquement justifiée, risquent d'être difficilement explicables.

Le changement des signes d'un vecteur latent (si le signe d'un élément est changé, il faut les changer tous) ne modifie pas les propriétés algébriques de la matrice des saturations. Cependant l'interprétation doit être modifiée; il faut considérer le facteur correspondant comme inversé dans son orientation: par exemple introversion plutôt qu'extroversion, difficulté plutôt que facilité, etc.

L'absence chez certains sujets d'un ou de plusieurs scores constitue, comme dans toute étude inférentielle, un problème sérieux. Tous les résultats sont en effet nécessaires au calcul des corrélations. S'il n'y a pas lieu de croire à des absences intentionnelles, on peut se permettre d'ignorer les sujets qui en sont affectés.

Le nombre de sujets sur lequel on effectue une analyse factorielle est de toute première importance. Une étude sur k variables donne lieu à $(k^2 - k)/2$ corrélations des variables prise deux à deux: ce qui constitue la matrice de corrélations. La probabilité d'apparition d'une corrélation significative, pour la population dont la corrélation est nulle, grandit avec le nombre de corrélations calculées et la petitesse de l'échantillon. Dans le cas d'un petit échantillon, la matrice de corrélations pourrait facilement, par hasard, être truffée de corrélations très erronées: ce qui aurait pour effet de révéler une structure factorielle fausse. Nunnally recommande en pratique de ne point travailler sur un échantillon qui soit plus petit que dix fois le nombre de variables.

L'utilisation de grosses batteries de variables, par exemple les questionnaires interminables, comporte de nombreux dangers. Obéissant au désir de tout savoir le chercheur est tenté de multiplier les variables, sans connaissance raisonnable de leur nature, ce qui conduit à des structures inexplicables. On a vu dans l'article précédent combien l'interprétation est dépendante de la connaissance interne de chaque variable.

Il ne faut pas utiliser l'analyse factorielle en n'importe quelle circonstance. C'est un instrument qui aide à mettre en évidence des structures: il est donc essentiel que les variables retenues appartiennent à un

système. L'intérêt serait faible de soumettre à cette technique un amas de variables. Les facilités qu'offre aujourd'hui l'ordinateur sont loin de diminuer un tel risque. Des phénomènes aussi variés que les poussées inflationnistes, les chasses aux sorcières, les succès des mouvements évangélistes, soumis à l'analyse factorielle feraient-ils apparaître comme explication commune les oppositions idéologiques des grands blocs politiques?

---

[1]Vedrine J.-P. "Le traitement des données en marketing", Ed. Organisation, Paris, 1991.

[2]Hubert Laforge "Analyse multivariée pour les sciences sociales et biologiques avec applications des logiciels BMP, BMDP, SPSS, SAS", Ed. Etudes Vivantes, Montréal, 1981.

[3]Le tableux est annexé à la fin du document

[4]adapté d'après M. Delattre dans J-.P. Lambin "La recherche marketing", McGraw - Hill, Paris, 1990 p.306.