Analyse multidimensionnelle appliquée

Denis Larocque, Léo Belzile

Version du 2020-01-24

Table des matières

Chapitre 1

Introduction

1.1 Survol du cours

1.1.1 Analyse factorielle exploratoire

On dispose de p variables $X_1, \ldots, X_p$. Peut-on expliquer les interrelations (la structure de corrélation) entre ces variables à l'aide d'un certain nombre (moins de p) de facteurs latents (non observés) ?

L'analyse factorielle est souvent utilisée pour analyser des questionnaires (construction d'échelles) comme dans l'exemple suivant.

Exemple 1.1. Pour les besoins d'une enquête, on a demandé à 200 consommateurs adultes de répondre aux questions suivantes par rapport à un certain type de magasin :

Sur une échelle de 1 à 5,

1. pas important
2. peu important
3. moyennement important
4. assez important
5. très important

Pour vous, à quel point est-ce important...

1. que le magasin offre de bons prix tous les jours ?
2. que le magasin accepte les cartes de crédit majeures (Visa, Mastercard) ?
3. que le magasin offre des produits de qualité ?
4. que les vendeurs connaissent bien les produits ?
5. qu'il y ait des ventes spéciales régulièrement ?
6. que les marques connues soient disponibles ?
7. que le magasin ait sa propre carte de crédit ?
8. que le service soit rapide ?
9. qu'il y ait une vaste sélection de produits ?

10. que le magasin accepte le paiement par carte de débit?
11. que le personnel soit courtois?
12. que le magasin ait en stock les produits annoncés?

Pouvons-nous identifier un nombre restreint de facteurs (concepts, dimensions) qui pourraient bien rendre compte de la structure de corrélation entre ces 12 variables?

Buts :

— Décrire et comprendre la structure de corrélation d'un ensemble de variables à l'aide d'un nombre restreint de concepts (appelés facteurs).
— Réduire le nombre de variables en créant une nouvelle variable par facteur. Ces nouvelles variables pourront par la suite être utilisées dans d'autres analyses (régression linéaire multiple par exemple).

1.1.2 Sélection de variables et de modèles

Dans plusieurs situations, on doit développer un modèle de prévision. Par exemple, on pourrait devoir développer un modèle pour :

— Détecter les faillites des clients (ou des entreprises)
— Cibler les clients qui seront intéressés par une offre promotionnelle
— Détecter les fraudes (par carte de crédit ou dans les rapports de revenus)
— Prévoir si un client va nous quitter.

Il y a en général plusieurs variables explicatives potentielles, et aussi plusieurs types de modèles possibles (régression linéaire, réseaux de neurones, arbres de régression ou de classification, etc.). Dans ce chapitre, nous verrons des principes généraux et des outils afin de sélectionner des modèles performants, ou bien un sous-ensemble de variables avec un bon pouvoir prévisionnel.

1.1.3 Régression logistique

On cherche à expliquer le comportement d'une variable binaire Y $(0-1)$, à l'aide de p variables quelconques $X_1,\ldots,X_p$.

Exemple 1.2. Une banque offre aux gens la possibilité de faire une demande de carte de crédit en ligne en promettant une approbation (conditionnelle) en quelques minutes seulement. Le tout est basé sur un modèle automatique de classification qui décide d'accorder ou non la carte ($Y=1$ ou $Y=0$) en fonction des réponses fournies par les clients potentiels à différentes questions comme : quel est votre revenu annuel brut (X_1), avez-vous d'autres cartes de crédit (X_2), êtes-vous locataire ou propriétaire (X_3), etc...

Buts :

— Comprendre comment et dans quelle mesure les variables $\boldsymbol{X}$ influencent la catégorie d'appartenance de Y.
— Développer un modèle pour faire de la classification, c'est-à-dire, prévoir la catégorie d'appartenance de Y pour un nouveau sujet à partir des variables $\boldsymbol{X}$.

1.1.4 Analyse de regroupements

On cherche à créer des groupes (« *clusters* ») d'individus homogènes en utilisant p variables $X_1,\ldots,X_p$.

Exemple 1.3. Cette méthode est utilisée en marketing pour la **segmentation de marché**, qui consiste en

> ...définir des sous-groupes réunissant des consommateurs qui partagent les mêmes préférences ou qui réagissent de façon semblable à des variables de marketing [1]

But :

— Combiner des sujets en groupes (interprétables) de telle sorte que les individus d'un même groupe soient les plus semblables possible par rapport à certaines caractéristiques et que les groupes soient les plus différents possible.

1.1.5 Analyse de survie

On s'intéresse au temps avant qu'un événement survienne. Par exemple :

— Temps qu'un client demeure abonné à un service offert par notre compagnie.
— Temps de survie d'un individu après avoir été diagnostiqué avec un certain type de cancer.
— Temps qu'un employé demeure au service de la compagnie.
— Temps qu'une franchise demeure en activité.
— Temps avant la faillite d'une entreprise (ou d'un particulier).
— Temps avant le prochain achat d'un client.

On observe chaque sujet jusqu'à ce que l'une des deux choses suivantes se produise : l'événement survient avant la fin de la période d'observation ou bien l'étude se termine et l'événement n'est toujours pas survenu. Dans le premier exemple, l'événement correspond au fait d'interrompre son abonnement. On dispose donc d'une variable temps T pour chaque individu qui est soit censurée, soit non censurée. Si l'individu a expérimenté l'événement avant la fin de la période d'observation, la valeur de est non censurée. Si l'événement n'est toujours pas survenu à la fin de la période d'observation, la valeur de T est censurée. Pour chaque individu, on dispose également d'un ensemble de variables explicatives $X_1, \ldots, X_p$.

But :

— Étudier les effets des variables explicatives sur le temps de survie et obtenir des prévisions du temps de survie.

1.1.6 Données manquantes

Il arrive fréquemment d'avoir des valeurs manquantes dans notre échantillon.

Simplement ignorer les sujets avec des valeurs manquantes et faire l'analyse avec les autres sujets conduit généralement à des estimations biaisées et à de l'inférence invalide.

Dans ce chapitre, nous verrons une méthode très générale afin de traiter les données manquantes, l'imputation multiple. Nous verrons comment elle peut être utilisée dans un contexte d'inférence et dans un contexte de prévision.

1. d'Astous, A. (2000). *Le projet de recherche en marketing*, 2e édition. Chenelière/McGraw-Hill.

Chapitre 2

Analyse factorielle exploratoire

2.1 Introduction

On dispose de p variables $X_1, \ldots, X_p$.

— Y a-t-il des groupements de variables ?
— Est-ce que les variables faisant partie d'un groupement semblent mesurer certains aspects d'un facteur commun (non observé) ?

Un tel groupement peut être détecté si plusieurs variables sont très corrélées entre elles. Est-ce que la structure de corrélation entre les p variables peut être expliquée à l'aide d'un nombre restreint de facteurs ?

Exemple de facteurs : Habileté quantitative, habileté sociale, importance accordée à la qualité du service, importance accordée à la loyauté, habileté de leader, etc…

L'analyse factorielle est aussi une méthode de réduction du nombre de variables. En effet, une fois qu'on a identifié les facteurs, on peut remplacer les variables individuelles par un résumé pour chaque facteur (qui est souvent la moyenne des variables qui font partie du facteur).

Pour faire une analyse factorielle, la taille d'échantillon devrait être d'au moins 10 fois le nombre de variables.

2.2 Rappels sur le coefficient de corrélation linéaire

On veut examiner la relation entre deux variables X_j et X_k et on dispose de n couples d'observations, où $x_{i,j}$ (respectivement $x_{i,k}$) est la valeur de la variable X_j (X_k) pour le ie individu.

Le coefficient de corrélation linéaire entre X_j et X_k, que l'on note $r_{j,k}$, cherche à mesurer la force de la relation linéaire entre deux variables, c'est-à-dire à quantifier à quel point les observations sont alignées autour d'une droite. Le coefficient de corrélation est

$$r_{j,k} = \frac{\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)(x_{i,k} - \overline{x}_k)}{\left\{\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)^2 \sum_{i=1}^{n}(x_{i,k} - \overline{x}_k)^2\right\}^{1/2}}$$

Les propriétés les plus importantes du coefficient de corrélation linéaire r sont les suivantes :

1) $-1 \leq r \leq 1$;
2) $r = 1$ (respectivement $r = -1$) si et seulement si les n observations sont exactement alignées sur une droite de pente positive (négative). C'est-à-dire, s'il existe deux constantes a et $b > 0$ ($b < 0$) telles que $y_i = a + bx_i$ pour tout $i = 1, \ldots, n$.

Règle générale,

— Plus la corrélation est près de 1, plus les points auront tendance à être alignés autour d'une droite de pente positive. Par conséquent, plus la valeur de X augmente, plus celle de Y aura tendance à augmenter et vice-versa.
— Plus la corrélation est près de -1, plus les points auront tendance à être alignés autour d'une droite de pente négative. Par conséquent, plus la valeur de X augmente, plus celle de Y aura tendance à diminuer et vice-versa.
— Lorsque la corrélation est presque nulle, les points n'auront pas tendance à être alignés autour d'une droite. Il est très important de noter que cela n'implique pas qu'il n'y a pas de relation entre les deux variables. Cela implique seulement qu'il n'y a pas de **relation linéaire** entre les deux variables.

2.3 Exemple de questionnaire

Le questionnaire suivant porte sur une étude dans un magasin. Pour les besoins d'une enquête, on a demandé à 200 consommateurs adultes de répondre aux questions suivantes par rapport à un certain type de magasin sur une échelle de 1 à 5, où

1. pas important
2. peu important
3. moyennement important
4. assez important
5. très important

Pour vous, à quel point est-ce important…

1. que le magasin offre de bons prix tous les jours ?
2. que le magasin accepte les cartes de crédit majeures (Visa, Mastercard) ?
3. que le magasin offre des produits de qualité ?
4. que les vendeurs connaissent bien les produits ?
5. qu'il y ait des ventes spéciales régulièrement ?
6. que les marques connues soient disponibles ?
7. que le magasin ait sa propre carte de crédit ?
8. que le service soit rapide ?
9. qu'il y ait une vaste sélection de produits ?
10. que le magasin accepte le paiement par carte de débit ?
11. que le personnel soit courtois ?
12. que le magasin ait en stock les produits annoncés ?

Une analyse factorielle cherchera à identifier automatiquement des groupes de variables qui sont fortement corrélées entre elles.

Les commandes **SAS** (ainsi que plusieurs commentaires) pour faire les analyses se trouvent dans le fichier `MATH60602_cours3.sas`. Les statistiques descriptives ainsi que la matrice des corrélations sont obtenues en exécutant les lignes suivantes :

```
proc corr data=multi.factor2;
var x1-x12;
run;
```

Statistiques simples

Variable	N	Moyenne	Ec-type	Somme	Minimum	Maximum	Libellé
x1	200	2.26	1.13	451	1	5	x1
x2	200	2.51	1.24	502	1	5	x2
x3	200	3.01	1.19	601	1	5	x3
x4	200	2.91	1.33	582	1	5	x4
x5	200	3.55	1.17	710	1	5	x5
x6	200	2.14	1.14	428	1	5	x6
x7	200	1.82	1.06	364	1	5	x7
x8	200	2.92	1.32	583	1	5	x8
x9	200	3.04	1.12	608	1	5	x9
x10	200	2.59	1.32	518	1	5	x10
x11	200	2.99	1.33	597	1	5	x11
x12	200	3.45	1.16	690	1	5	x12

Coefficients de corrélation de Pearson, N = 200
Proba > |r| sous H0: Rho=0

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
x1	1.00	-0.08	-0.14	-0.07	0.38	-0.01	-0.10	-0.13	-0.03	-0.11	-0.12	-0.01
x1		0.25	0.06	0.34	0.00	0.91	0.18	0.06	0.70	0.12	0.09	0.87
x2	-0.08	1.00	0.04	-0.02	-0.08	0.06	0.50	0.01	-0.01	0.43	-0.12	0.07
x2	0.25		0.55	0.80	0.26	0.43	0.00	0.84	0.92	0.00	0.09	0.35
x3	-0.14	0.04	1.00	0.10	-0.06	0.39	0.00	0.05	0.47	0.08	0.13	0.46
x3	0.06	0.55		0.15	0.43	0.00	0.99	0.53	0.00	0.29	0.07	0.00
x4	-0.07	-0.02	0.10	1.00	-0.05	0.06	0.08	0.57	0.01	0.09	0.50	0.09
x4	0.34	0.80	0.15		0.52	0.39	0.25	0.00	0.86	0.22	0.00	0.22
x5	0.38	-0.08	-0.06	-0.05	1.00	-0.04	-0.04	-0.02	0.03	-0.07	-0.06	-0.07
x5	0.00	0.26	0.43	0.52		0.58	0.60	0.83	0.64	0.34	0.43	0.34
x6	-0.01	0.06	0.39	0.06	-0.04	1.00	0.07	0.04	0.32	0.07	-0.04	0.32
x6	0.91	0.43	0.00	0.39	0.58		0.32	0.56	0.00	0.36	0.56	0.00
x7	-0.10	0.50	0.00	0.08	-0.04	0.07	1.00	0.09	-0.02	0.51	-0.03	0.02
x7	0.18	0.00	0.99	0.25	0.60	0.32		0.22	0.74	0.00	0.63	0.76
x8	-0.13	0.01	0.05	0.57	-0.02	0.04	0.09	1.00	-0.03	0.16	0.55	0.04
x8	0.06	0.84	0.53	0.00	0.83	0.56	0.22		0.62	0.02	0.00	0.53
x9	-0.03	-0.01	0.47	0.01	0.03	0.32	-0.02	-0.03	1.00	0.01	0.02	0.39
x9	0.70	0.92	0.00	0.86	0.64	0.00	0.74	0.62		0.91	0.77	0.00
x10	-0.11	0.43	0.08	0.09	-0.07	0.07	0.51	0.16	0.01	1.00	0.01	0.02
x10	0.12	0.00	0.29	0.22	0.34	0.36	0.00	0.02	0.91		0.91	0.75
x11	-0.12	-0.12	0.13	0.50	-0.06	-0.04	-0.03	0.55	0.02	0.01	1.00	0.05
x11	0.09	0.09	0.07	0.00	0.43	0.56	0.63	0.00	0.77	0.91		0.48
x12	-0.01	0.07	0.46	0.09	-0.07	0.32	0.02	0.04	0.39	0.02	0.05	1.00
x12	0.87	0.35	0.00	0.22	0.34	0.00	0.76	0.53	0.00	0.75	0.48	

2.4 Description du modèle d'analyse factorielle

On dispose d'observations sur p variables $X_1, \ldots, X_p$. Le modèle d'analyse factorielle fait l'hypothèse que ces variables dépendent linéairement d'un plus petit nombre m de variables aléatoires, $F_1, \ldots, F_m$, appelées facteurs communs et de p termes d'erreurs (ou facteurs spécifiques) $\varepsilon_1, \ldots, \varepsilon_p$, de moyenne $\mathsf{E}(\varepsilon_i) = 0$ et de

variance $\mathsf{Var}(\varepsilon_i) = \psi_i$ pour $i = 1, \ldots, p$. Spécifiquement, le modèle est

$$\begin{aligned} X_1 &= \mu_1 + l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1 \\ X_2 &= \mu_2 + l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p &= \mu_p + l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p, \end{aligned}$$

où μ_i est l'espérance de la variable aléatoire X_i ($i = 1, \ldots, p$) et où l_{ij} est le chargement de la variable X_i sur le facteur F_j ($i = 1, \ldots, p$; $j = 1, \ldots, m$).

Les espérances (μ_i), les chargements (l_{ij}) et les variances (ψ_i) sont des quantités fixes, mais inconnues, tandis que les facteurs communs (F_i) et spécifiques (ε_i) sont des variables aléatoires non observables que l'on assume non corrélée aux facteurs F et entre elles.

Des hypothèses supplémentaires sont nécessaires afin de pouvoir utiliser ce modèle (contraintes d'identifiabilité des paramètres). Sans entrer dans les détails, mentionnons que l'une de ces hypothèses est que les facteurs sont non corrélés.

De plus, si les variables ont été préalablement standardisées de telle sorte que $\mathsf{E}(X_i) = 0$ et $\mathsf{Var}(X_i) = 1$ (note : ceci revient à utiliser la matrice de corrélation des observations dans l'analyse ce qui est fait par défaut dans **SAS**), alors $\mathsf{Cor}\left(X_i, F_j\right) = l_{ij}$, c'est-à-dire, le chargement de la variable X_i sur le chargement F_j est le coefficient de corrélation entre cette variable et ce facteur.

Sans aucune contrainte sur le modèle, la matrice de covariance de $X_1, \ldots, X_p$ possède $p(p+1)/2$ paramètres, soit p variances et $p(p-1)/2$ termes de corrélation. Avec le modèle d'analyse factorielle, on suppose que l'on peut décrire cette structure en utilisant seulement $p(m+1)$ paramètres (p variances spécifiques et pm chargements). Par exemple, avec $p = 50$ variables et $m = 6$ facteurs, on essaie de décrire la structure de covariance à l'aide de 350 paramètres au lieu de 1275.

Il existe plusieurs méthodes pour extraire les facteurs, c'est-à-dire pour estimer les paramètres du modèle (les ψ_i et les l_{ij}). Nous allons discuter de deux d'entre elles : la méthode du maximum de vraisemblance et la méthode des composantes principales. L'avantage de l'estimation par maximum de vraisemblance est qu'elle permet l'utilisation de critères d'information et de statistiques de tests pour guider le choix du nombre de facteurs. En revanche, l'estimation des paramètres requiert une optimisation numérique qui peut être délicate selon les cas de figure.

2.4.1 Rotation des facteurs

Dans le modèle d'analyse factorielle, on peut montrer que, lorsqu'il y a deux facteurs ou plus, il existe plusieurs configurations de facteurs qui donnent la même structure de covariance. En fait, les chargements peuvent seulement être déterminés à une transformation orthogonale prêt (note : une transformation orthogonale est une transformation qui préserve le produit scalaire ; elle préserve ainsi toutes les distances et les angles entre deux vecteurs). Si les chargements provenant d'une méthode d'extraction des facteurs ne sont pas uniques, la matrice de corrélation estimée par le modèle est par contre unique.

Il existe plusieurs techniques de rotation de facteurs. Le but de ces techniques est d'essayer de trouver une solution qui fera en sorte que les facteurs seront facilement interprétables. La méthode la plus utilisée est la méthode **varimax** : elle produit une configuration de chargement en maximisant la variance de la somme

des carrés des chargements pour les m facteurs. La méthode varimax tend à produire une configuration de facteurs tel que les chargements de chaque variable sont dispersés (des chargements élevés positifs ou négatifs et d'autres presque nuls).

Je vous suggère de toujours tenter d'interpréter la solution avec une rotation varimax. Si ce n'est pas suffisamment clair, il existe d'autres méthodes de rotation dont certaines (les rotations de type oblique) permettent la présence de corrélation entre les facteurs.

2.5 Estimation des facteurs

Les chargements estimés pour la solution à quatre facteurs, suite à la rotation varimax, sont obtenus avec le code SAS suivant :

```
proc factor data=multi.factor2
 method=ml rotate=varimax nfact=4
 maxiter=500 flag=.3 hey;
 var x1-x12;
run;
```

Caractéristique du facteur de rotation

		Factor1	Factor2	Factor3	Factor4
x1	x1	-8	-2	-6	99 *
x2	x2	-8	5	67 *	-5
x3	x3	8	75 *	1	-11
x4	x4	71 *	7	6	0
x5	x5	-2	-3	-5	37 *
x6	x6	1	51 *	8	1
x7	x7	3	0	75 *	-5
x8	x8	79 *	-1	10	-6
x9	x9	-3	63 *	-4	-2
x10	x10	10	5	66 *	-6
x11	x11	71 *	5	-10	-7
x12	x12	5	61 *	3	1

Les valeurs imprimées sont multipliées par 100 et arrondies au nombre entier le plus proche.
Les valeurs supérieures à 0.3 sont indiquées par un signe '*'.

En général, on associe une variable à un groupe (facteur) si son chargement est supérieur à 0, 3 (en valeur absolue), ce qui donne

— Facteur 1 : X_4, X_8 et X_{11}
— Facteur 2 : X_3, X_6, X_9 et X_{12}
— Facteur 3 : X_2, X_7 et X_{10}
— Facteur 4 : X_1 et X_5.

Ces facteurs sont interprétables :

— Le facteur 1 représente l'importance accordée au service.
— Le facteur 2 représente l'importance accordée aux produits.
— Le facteur 3 représente l'importance accordée à la facilité de paiement.
— Le facteur 4 représente l'importance accordée aux prix.

Dans cet exemple, les choses se sont bien passées et le nombre de facteurs que nous avons spécifié (4) semble être adéquat, mais ce n'est pas toujours aussi évident. Il est utile d'avoir des outils pour guider le choix du nombre de facteurs.

2.6 Choix du nombre de facteurs

Il existe différentes méthodes pour se guider dans le nombre de facteurs, m, à utiliser. Cependant, le point important à retenir est que, peu importe le nombre choisi, il faut que les facteurs soient **interprétables**. Par conséquent, les méthodes qui suivent ne devraient servir que de guide et non pas être suivies aveuglément. La méthode du maximum de vraisemblance que nous avons utilisée dans l'exemple possède l'avantage de fournir trois critères pour choisir le nombre de facteurs appropriés. Ces critères sont :

— AIC (critère d'information d'Akaike)
— BIC (critère d'information bayésien de Schwarz)
— Le test du rapport de vraisemblance pour l'hypothèse nulle que le modèle de corrélation décrit le modèle factoriel avec m facteurs est adéquat, contre l'alternative qu'il n'est pas adéquat.

Les critères d'information servent à la sélection de modèles ; ils seront traités plus en détail dans les chapitres qui suivent. Pour l'instant, il est suffisant de savoir que le modèle avec la valeur du critère AIC (ou BIC) la plus petite est considéré le « meilleur » (selon ce critère).

Les sorties suivantes proviennent du même programme SAS et correspondent au modèle factoriel avec quatre facteurs estimé par maximum de vraisemblance.

Tests de significativité basés sur 200 observations

Test	DDL	Khi-2	Pr > khi-2
H0: Aucun facteur commun	66	503.4490	<.0001
HA: Au moins un facteur commun			
H0: 4 facteurs suffisants	24	12.5708	0.9727
HA: davantage de facteurs sont requis			

Khi-2 sans correction de Bartlett	13.06317
Critère d'information d'Akaike	-34.93683
Critère bayésien de Schwarz	-114.09645
Coefficient de fiabilité de Tucker et Lewis	1.07185

Pour choisir le nombre de facteurs avec les critères d'information, il faut ajuster le modèle en faisant varier le nombre de facteurs (option `nfact`) et extraire la valeur numérique correspondante.

Le tableau 2.1 présente les valeurs estimées des critères d'information et des valeurs-p pour le test du rapport de vraisemblance pour cinq modèles. Le critère AIC suggère quatre facteurs, tandis que les deux autres critères (BIC et test du rapport de vraisemblance suggèrent plutôt trois facteurs.

TABLE 2.1: Critères d'information et valeurs-p pour le modèle factoriel à m facteurs

m	AIC	BIC	valeur-p
1	228, 0	49, 9	<0, 001
2	99, 5	-42, 3	<0, 001
3	-20, 5	-129, 3	0, 096
4	-34, 9	-114, 1	0, 973
5	-24, 8	-77, 6	0, 975

On peut considérer le modèle avec trois facteurs : les chargements (après rotation varimax) sont données dans la figure 2.1.

Cette solution récupère les trois facteurs *service*, *produits* et *paiement* de la solution précédente à quatre facteurs. Le facteur *prix* (qui était formé de X_1 et X_5) n'est plus présent : que faire avec ce dernier? Cela dépend du but de l'analyse et nous y reviendrons plus tard.

Pour terminer cette section, voici la description de deux autres critères *classiques* pour choisir le nombre de facteurs. Ces deux critères sont :

— Critère de Kaiser, un critère basé sur les valeurs propres. Avec une analyse en composantes principales basée sur la matrice des corrélations, la valeur propre associée à un facteur représente la partie de

		Factor1	Factor2	Factor3
x1	**x1**	-15	-9	-14
x2	**x2**	-9	3	67 *
x3	**x3**	10	76 *	4
x4	**x4**	71 *	5	7
x5	**x5**	-5	-6	-10
x6	**x6**	1	50 *	9
x7	**x7**	2	-3	75 *
x8	**x8**	79 *	-3	12
x9	**x9**	-2	63 *	-2
x10	**x10**	9	3	67 *
x11	**x11**	72 *	5	-8
x12	**x12**	6	60 *	5

Les valeurs imprimées sont multipliées par 100 et arrondies au nombre entier le plus proche.
Les valeurs supérieures à 0.3 sont indiquées par un signe '*'.

FIGURE 2.1 – Estimés des chargements pour trois facteurs avec rotation varimax

la variance totale qui est expliquée par ce facteur. Chaque variable compte pour un dans la variance totale. Le nombre de facteurs choisis est le nombre de valeurs propres supérieures à 1. L'idée est de garder seulement les facteurs qui expliquent plus de variance qu'une variable individuelle.

— le diagramme d'éboulis : un graphique des valeurs propres ordonnées de la plus grande à la plus petite en fonction de $1,\ldots,p$. Habituellement, ce graphe prendra la forme d'une chute assez importante suivie d'une stabilisation des valeurs propres. Avec ce critère, le nombre de facteurs est déterminé par le nombre de valeurs propres avant le début du coude où il a stabilisation apparente. L'idée est de choisir l'endroit où l'ajout d'un facteur supplémentaire n'apporte qu'un gain marginal faible. Ce critère est par contre subjectif et dépend de l'analyste. En ajoutant `scree` comme option à `proc factor`, on obtient le diagramme d'éboulis mais il est facile de le créer manuellement et le résultat est esthétiquement plus réussi.

Les sorties qui suivent proviennent du programme :

```
proc factor data=multi.factor2 method=principal
 scree rotate=varimax flag=.3;
 ods output Eigenvalues=eigen;
 var x1-x12;
run;

proc sgplot data=eigen;
 scatter x=number y=eigenvalue;
 yaxis label="valeurs propres";
 xaxis label='nombre';
run;
```

Cette fois-ci, c'est la méthode des composantes principales qui est utilisée ; cette dernière consiste à estimer les chargements en utilisant les m premières valeurs propres et vecteurs propres de la matrice de corrélation. En ne spécifiant pas l'option `nfact`, **SAS** choisit le nombre de facteurs en utilisant par défaut le critère de Kaiser (valeurs propres supérieures à 1). Quatre facteurs sont retenus, tel qu'indiqué par la sortie au bas du tableau 2.2. Pour le diagramme d'éboulis de la figure 2.3, le choix est assez subjectif : il semble raisonnable de choisir trois ou quatre facteurs.

On suggère d'utiliser *de facto* les trois critères découlant de l'utilisation de la vraisemblance et de déterminer le nombre de facteurs à extraire selon différents critères avant d'examiner les modèles avec ce nombre de facteurs et ceux avec un facteur de moins ou de plus. Au final, le plus important est de pouvoir interpréter raisonnablement les facteurs et donc le modèle retenu est souvent choisi selon le critère **Wow!**. On veut dire par là que la configuration de facteurs choisie est compréhensible.

2.7 Construction d'échelles à partir des facteurs

Si le seul but de l'analyse factorielle est de comprendre la structure de corrélation entre les variables, alors se limiter à l'interprétation des facteurs est suffisant.

Si par contre, le but est de réduire le nombre de variables pour pouvoir par la suite procéder à d'autres analyses statistiques, l'analyse factorielle peut alors servir de guide pour construire de nouvelles variables (échelles). En supposant que l'analyse factorielle a produit des facteurs qui sont interprétables et satisfaisants, la méthode de construction d'échelles la plus couramment utilisée consiste à construire m nouvelles variables,

Valeurs propres de la matrice de corrélation: Total = 12 Moyenne = 1

	Valeur propre	Différence	Proportion	Cumulé
1	2.42730801	0.42845477	0.2023	0.2023
2	1.99885324	0.05649946	0.1666	0.3688
3	1.94235378	0.64106103	0.1619	0.5307
4	1.30129275	0.56297464	0.1084	0.6392
5	0.73831811	0.04657350	0.0615	0.7007
6	0.69174461	0.12512145	0.0576	0.7583
7	0.56662316	0.02697168	0.0472	0.8055
8	0.53965148	0.03002176	0.0450	0.8505
9	0.50962972	0.03638282	0.0425	0.8930
10	0.47324690	0.01804874	0.0394	0.9324
11	0.45519816	0.09941808	0.0379	0.9704
12	0.35578007		0.0296	1.0000

4 facteurs seront retenus par le critère MINEIGEN.

FIGURE 2.2 – Valeurs propres et proportion de variance

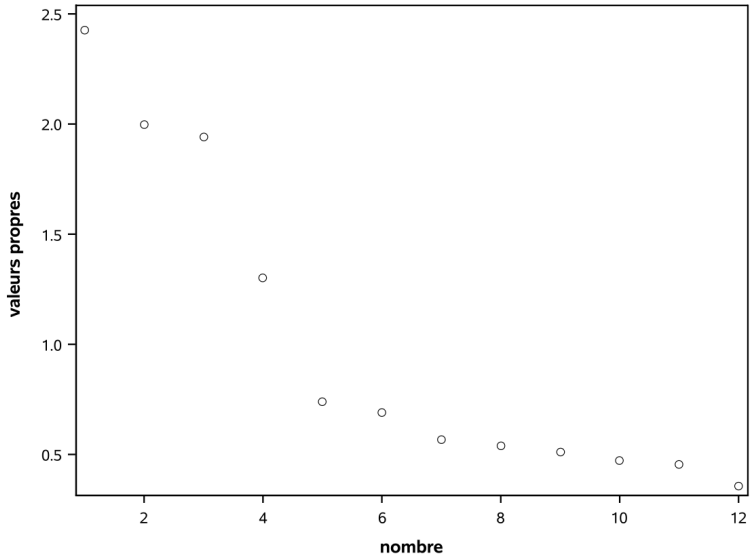

FIGURE 2.3 – Diagramme d'éboulis

une par facteur. Pour un facteur donné, la nouvelle variable est simplement la moyenne des variables ayant des chargements élevés sur ce facteur (positifs ou négatifs, mais de même signe). Une autre méthode, les scores factoriels, sera présentée plus loin.

Lorsqu'on construit une échelle, il est important d'examiner sa cohérence interne. Ceci peut être fait à l'aide du coefficient alpha de Cronbach. Ce coefficient mesure à quel point chaque variable faisant partie d'une échelle est corrélée avec le total de toutes les variables pour cette échelle. Plus le coefficient est élevé, plus les variables ont tendance à être corrélées entre elles. L'alpha de Cronbach est

$$\alpha = \frac{k}{k-1}\frac{S^2 - \sum_{i=1}^{k} S_i^k}{S^2},$$

où k est le nombre de variables dans l'échelle, S^2 est la variance empirique de la somme des variables et S_i^2 est la variance empirique de la ie variable. En pratique, on voudra que ce coefficient soit au moins égal à 0, 6 pour être satisfait de la cohérence interne de l'échelle.

Avec **SAS**, la procédure `corr` permet de calculer α.

```
/* pour le facteur service */
proc corr data=multi.factor2 alpha;
var x4 x8 x11;
run;
/* pour le facteur produits */
proc corr data=multi.factor2 alpha;
var x3 x6 x9 x12;
run;
/* pour le facteur paiement */
```

```
proc corr data=multi.factor2 alpha;
var x2 x7 x10;
run;
/* pour le facteur prix */
proc corr data=multi.factor2 alpha;
var x1 x5;
run;
```

Coefficient Alpha de Cronbach

Variables	Alpha
Brut	0.780524
Normalisé	0.780611

Coefficient Alpha de Cronbach avec variable supprimée

	Variables brutes		Variables standardisées		
Variable supprimée	Corrélation avec total	Alpha	Corrélation avec total	Alpha	Libellé
x4	0.609256	0.712551	0.609409	0.712565	x4
x8	0.649038	0.668928	0.649028	0.668929	x8
x11	0.595499	0.727412	0.595641	0.727434	x11

FIGURE 2.4 – Alpha de Cronbach pour le facteur *service*.

Il faut utiliser le alpha brut. Ainsi, les alphas de Cronbach sont tous satisfaisants (plus grand que 0, 6) sauf pour le facteur *prix* ($\alpha = 0,546$). **SAS** fournit également la matrice des corrélations des variables de l'échelle ainsi que la valeur du alpha de Cronbach si on retirait une variable à la fois de l'échelle. Tout est donc cohérent. Les échelles provenant des facteurs *service*, *produits* et *paiement*, sont satisfaisantes. Ces facteurs sont identifiés à la fois dans la solution à quatre, mais aussi dans la solution à troisfacteurs. Le facteur *prix* est celui qui apparaît en plus dans la solution à quatre facteurs. Il a une interprétation claire, mais son faible alpha ferait en sorte qu'il serait discutable de travailler avec l'échelle *prix* dans d'autres analyses (du moins avec selon l'usage habituel du alpha).

2.8 Compléments d'information

2.8.1 Variables ordinales

Théoriquement, une analyse factorielle ne devrait être faite qu'avec des variables continues. Par contre, en pratique, on l'utilise souvent aussi avec des variables ordinales (comme pour l'exemple portant sur le questionnaire) et même avec des variables binaires (0-1).

Dans ce genre de situation, on peut aussi utiliser d'autres mesures d'associations au lieu du coefficient

Coefficient Alpha de Cronbach	
Variables	**Alpha**
Brut	0.718253
Normalisé	0.717602

Coefficient Alpha de Cronbach avec variable supprimée					
	Variables brutes		Variables standardisées		
Variable supprimée	**Corrélation avec total**	**Alpha**	**Corrélation avec total**	**Alpha**	**Libellé**
x3	0.584387	0.606777	0.584428	0.606820	x3
x6	0.430389	0.699956	0.429825	0.699753	x6
x9	0.509722	0.654325	0.508628	0.653545	x9
x12	0.501808	0.658864	0.500831	0.658223	x12

FIGURE 2.5 – Alpha de Cronbach pour le facteur *produits*.

Coefficient Alpha de Cronbach	
Variables	**Alpha**
Brut	0.727492
Normalisé	0.734783

Coefficient Alpha de Cronbach avec variable supprimée					
	Variables brutes		Variables standardisées		
Variable supprimée	**Corrélation avec total**	**Alpha**	**Corrélation avec total**	**Alpha**	**Libellé**
x2	0.532476	0.661213	0.538157	0.672224	x2
x7	0.596466	0.601795	0.596509	0.602521	x7
x10	0.536537	0.663250	0.540698	0.669254	x10

FIGURE 2.6 – Alpha de Cronbach pour le facteur *paiement*.

Coefficient Alpha de Cronbach	
Variables	**Alpha**
Brut	0.545634
Normalisé	0.545805

FIGURE 2.7 – Alpha de Cronbach pour le facteur *prix*.

de corrélation linéaire. Par exemple, on peut utiliser la corrélation polychorique, qui est une mesure de corrélation entre deux variables ordinales. La corrélation tétrachorique correspond au cas spécial de deux variables binaires.

Ma suggestion est d'utiliser la corrélation linéaire ordinaire avec des variables ordinales (même binaires). Si les résultats ne sont pas satisfaisants, on peut alors essayer avec d'autres mesures d'associations.

On peut refaire l'analyse des données portant sur le magasin dans **SAS** en utilisant la corrélation polychorique calculées par la procédure `corr` et en passant la sortie à la procédure `factor`.

```
proc corr data=multi.factor2 polychoric out=poly_corr;
var x1-x12;
run;

proc factor data=poly_corr
 method=ml rotate=varimax nfact=4
 maxiter=500 flag=.3 hey;
 var x1-x12;
run;
```

Les chargements sont donnés dans la figure 2.8. Les facteurs obtenus sont les mêmes qu'en utilisant les corrélations linéaires.

2.8.2 Autres méthodes d'extractions de facteurs

Il n'y a pas de formule explicites pour l'estimation des paramètres avec la méthode du maximum de vraisemblance et un algorithme d'optimisation est nécessaire pour l'option des paramètres. Dans certains cas, l'algorithme peut terminer sans solution ou retourner un cas limite (où la variance est négative ou nulle). C'est le cas dans notre exemple avec quatre facteurs (solution de Heywood), bien que ce ne soit pas indiqué. La sortie **SAS** contient des informations sur la convergence de l'estimé : idéalement, on obtient la mention `Critère de convergence respecté`; autrement, essayez de varier le nombre. Un autre signe que l'algorithme n'a pas convergé est la présence de degrés de libertés négatifs pour le test du rapport de vraisemblance.

La méthode par les composantes principales (mentionnée lors de la présentation des valeurs propres et du diagramme d'éboulis a une solution explicite et peut donc dépanner si on n'arrive pas à obtenir le maximum de vraisemblance.

D'autres méthodes sont aussi disponibles dans **SAS** (voir la rubrique d'aide du logiciel) mais les deux

Caractéristique du facteur de rotation								
		Factor1		Factor2		Factor3		Factor4	
x1	**x1**	-8		-2		-6		99	*
x2	**x2**	-8		5		67	*	-5	
x3	**x3**	8		75	*	1		-11	
x4	**x4**	71	*	7		6		0	
x5	**x5**	-2		-3		-5		37	*
x6	**x6**	1		51	*	8		1	
x7	**x7**	3		0		75	*	-5	
x8	**x8**	79	*	-1		10		-6	
x9	**x9**	-3		63	*	-4		-2	
x10	**x10**	10		5		66	*	-6	
x11	**x11**	71	*	5		-10		-7	
x12	**x12**	5		61	*	3		1	

Les valeurs imprimées sont multipliées par 100 et arrondies au nombre entier le plus proche.
Les valeurs supérieures à 0.3 sont indiquées par un signe '*'.

FIGURE 2.8 – Chargements estimés pour la corrélation polychorique

méthodes mentionnées devraient être suffisantes pour la grande majorité des applications.

2.8.3 Autres méthodes de rotation des facteurs

Jusqu'à présent, nous avons utilisé la méthode de rotation orthogonale varimax. Il existe de nombreuses autres méthodes de rotations orthogonales telles, orthomax, quartimax, parsimax et equimax (voir la rubrique d'aide de **SAS**). Rappelez-vous que le modèle d'analyse factorielle de base suppose que les facteurs sont non corrélés. Les rotations de type obliques quant à elles permettent d'introduire de la corrélation entre les facteurs. Quelquefois, une telle rotation facilitera davantage l'interprétation des facteurs qu'une rotation orthogonale. **SAS** permet l'utilisation de plusieurs méthodes de rotation obliques qui sont documentées dans la rubrique d'aide. Notez qu'il faut être prudent lorsqu'on utilise une méthode de rotation oblique car il y aura trois matrices de chargements après rotation (coefficients de régression normalisés, corrélations semi-partielles ou corrélations). On suggère l'utilisation de la première, soit la représentation avec **coefficients de régression normalisés**. Il s'agit des coefficients de régression si on voulait prédire les variables à l'aide des facteurs. Ils indiquent donc à quel point chaque facteur est associé à chaque variable. Dans le cas d'une rotation orthogonale, ces trois matrices sont les mêmes et il s'agit de trois interprétations valides des chargements.

Le programme suivant fait une analyse factorielle avec quatre facteurs, mais en utilisant une rotation varimax oblique (option `rotate=obvarimax`).

```
proc factor data=multi.factor2
maxiter=500 flag=.3 hey;
var x1-x12;
run;
```

La matrice des corrélations entres facteurs est donnée dans la figure 2.9 et les chargements sont présentés dans la figure 2.10. On voit ici qu'on obtient les mêmes quatre facteurs qu'avec une rotation varimax orthogonale.

Corrélations inter-facteurs

	Factor1	Factor2	Factor3	Factor4
Factor1	100 *	7	4	-11
Factor2	7	100 *	6	-7
Factor3	4	6	100 *	-13
Factor4	-11	-7	-13	100 *

Les valeurs imprimées sont multipliées par 100 et arrondies au nombre entier le plus proche.
Les valeurs supérieures à 0.3 sont indiquées par un signe '*'.

FIGURE 2.9 – Corrélation interfacteurs pour rotation varimax oblique

Représentation du facteur avec rotation (Coefficients de régression normalisés)

		Factor1		Factor2		Factor3		Factor4	
x1	**x1**	-2		3		3		100	*
x2	**x2**	-9		2		67	*	-2	
x3	**x3**	6		75	*	-2		-10	
x4	**x4**	72	*	4		4		3	
x5	**x5**	0		-1		-2		37	*
x6	**x6**	0		51	*	7		2	
x7	**x7**	2		-3		75	*	-1	
x8	**x8**	79	*	-5		8		-3	
x9	**x9**	-4		63	*	-5		-1	
x10	**x10**	9		2		66	*	-3	
x11	**x11**	71	*	2		-11		-4	
x12	**x12**	4		61	*	2		2	

Les valeurs imprimées sont multipliées par 100 et arrondies au nombre entier le plus proche.
Les valeurs supérieures à 0.3 sont indiquées par un signe '*'.

FIGURE 2.10 – Chargements avec rotation oblique varimax

2.8.4 Scores factoriels

Avec les données de l'exemple, en nous basant sur les résultats de l'analyse factorielle, nous avons créé quatre nouvelles échelles (une par facteur) que l'on peut calculer pour chaque individu :

— *service* = $(X_4 + X_8 + X_{11})/3$,
— *produit* = $(X_3 + X_6 + X_9 + X_{12})/4$,
— *paiement* = $(X_2 + X_7 + X_{10})/3$,
— *prix* = $(X_1 + X_5)/2$.

Par exemple, la variable *prix* peut donc être vu comme une combinaison linéaire des 12 variables où seulement X_1 et X_5 reçoivent un poids (égal) différent de zéro. Une autre façon de créer de nouvelles variables consiste à calculer des scores factoriels (un pour chaque facteur) pour chaque individu. Par exemple, pour un individu i donné, le score factoriel pour le facteur k peut être prédit à l'aide de la formule

$$\begin{aligned}\hat{F}_{ik} &= \hat{\mathbf{L}}^\top \mathbf{R}^{-1} \boldsymbol{z} \\ &= \widehat{\gamma}_{1,k} z_{i,1} + \cdots + \widehat{\gamma}_{12,k} z_{i,12},\end{aligned}$$

où $z_{i,1}, \ldots, z_{i,12}$ sont les valeurs centrées et réduites des observations correspondant à l'individu et où $\widehat{\gamma}_{1,k}, \ldots, \widehat{\gamma}_{12,k}$ sont des coefficients estimés à partir des chargements l_{ij} (après rotation) et de la matrice de corrélation des variables $\mathbf{R}$, avec $\widehat{\gamma}_{i,k} = \sum_{j=1}^{p} \hat{l}_{kj} r_{jk}$.

Ainsi, chacune des 12 variables originales contribue au calcul du score factoriel. Les variables ayant des chargements plus élevés sur ce facteur auront tendance à avoir des poids ($\widehat{\gamma}$) plus élevés. Par contre, les scores factoriels ne sont pas uniques car ils dépendent des chargements utilisés (et donc à la fois de la méthode d'estimation et de la méthode de rotation). On peut également utiliser les scores factoriels au lieu des 12 variables originales dans des analyses subséquentes. Il est suggéré d'utiliser les nouvelles variables (échelles) obtenues en faisant les moyennes des variables identifiées comme faisant partie de chaque facteur pour les raisons suivantes :

— l'interprétation des scores factoriels est moins claire (chaque facteur dépend de toutes les variables)
— les scores factoriels ne sont pas uniques (ils dépendent de la méthode d'estimation et de rotation).
— les coefficients servant au calcul seront différents d'une étude à l'autre.

Pour obtenir les scores avec **SAS**, il suffit d'insérer l'option `score` à la procédure `factor`. L'option `out=...` permet de créer un fichier de données **SAS** qui contient la valeur des m scores pour chaque individu. Les scores factoriels pour l'exemples sont rapportés dans la figure 2.11. On remarque que :

— pour le premier facteur, trois variables ont des poids importants (X_4, X_8 et X_{11}). Il s'agit donc d'un facteur très proche du facteur *service*.
— pour le deuxième facteur, les variables X_3, X_6, X_9 et X_{12} ont des poids importants. Il s'agit donc d'un facteur très proche du facteur *produits*.
— pour le troisième facteur, les variables X_2, X_7, X_{10} ont des poids importants. Il s'agit donc d'un facteur très proche du facteur *paiement*.
— pour le quatrième facteur, seule la variable X_1 a un poids important. On aurait pu s'attendre à ce que ce soit également le cas pour X_5, en lien avec le facteur *prix* — ce facteur était moins clair selon le alpha de Cronbach.

Les corrélations entre les échelles (construites avec les moyennes) et les scores factoriels sont données dans la figure 2.12. On remarque la forte corrélation entre le score factoriel et les échelles construites avec les

Coefficients du score normalisés					
		Factor1	**Factor2**	**Factor3**	**Factor4**
x1	x1	0.03059	0.04866	0.04199	1.01161
x2	x2	-0.04524	0.01276	0.30701	0.01431
x3	x3	0.00898	0.45246	-0.01635	0.00975
x4	x4	0.30245	0.01130	0.01197	0.02657
x5	x5	0.00310	-0.00605	-0.00925	-0.00041
x6	x6	-0.00837	0.17490	0.02305	0.00447
x7	x7	-0.00542	-0.01815	0.44283	0.02490
x8	x8	0.45256	-0.04535	0.04422	0.03993
x9	x9	-0.02516	0.26576	-0.02469	0.00227
x10	x10	0.02061	0.00819	0.29821	0.01927
x11	x11	0.30260	0.00462	-0.06951	0.02170
x12	x12	0.00289	0.24472	0.00325	0.00581

FIGURE 2.11 – Coefficients du score normalisés

moyennes pour les facteurs *service, produits* et *paiement*. Cela veut dire qu'utiliser les échelles ou les scores factoriels ne devrait pas faire de différence dans des analyses subséquentes. Par contre, cette corrélation est plus faible (0.82) pour le facteur *prix*.

Coefficients de corrélation de Pearson, N = 200				
	Factor1	**Factor2**	**Factor3**	**Factor4**
service	0.99397	0.04659	0.02748	-0.05223
produit	0.04598	0.98350	0.03233	-0.03972
paiement	0.02640	0.04496	0.98615	-0.06790
prix	-0.07126	-0.03562	-0.07746	0.81920

FIGURE 2.12 – Corrélation entre scores et échelles