Université de Montréal

département de sociologie

L'analyse factorielle et l'analyse de fidélité

notes de cours et exemples

Notes aux lecteurs...

Ce texte a d'abord été préparé pour accompagner un cours d'introduction à l'analyse factorielle. J'ai tenté qu'il soit aussi complet que possible. Toutefois, la préoccupation première a évidemment influencé la forme.

Ce texte se veut une synthèse pratique visant à faciliter la compréhension, l'exécution et l'interprétation d'analyses factorielles et de fidélité dans le processus de la validation de mesures. Il s'inspire entre autres de la bibliographie présentée en fin de texte (particulièrement Tabachnik et Fidell, Using Multivariate Statistics) et de l'expérience de l'auteur dans la conduite de ce type d'analyse.

J'ai ajouté à la fin un texte de Lauri Tarkkonen qui explique bien à mon avis la différence entre analyse factorielle et analyse en composantes principales.

Je suis ouverte et intéressée à vos commentaires et critiques qui me permettront d'améliorer éventuellement le texte.

Vous pouvez consulter le site web du cours Méthodes quantitatives avancées pour en savoir plus et éventuellement écouter les présentations enregistrées relatives à l'analyse factorielle sur le site suivant:
http://www.mapageweb.umontreal.ca/durandc/menuMethodesQuantitatives.html#analyse_compo

Claire.Durand@umontreal.ca

Table des matières

1. Qu'est-ce que l'analyse factorielle? A quoi ça sert?

On utilise le terme générique d'analyse factorielle pour parler de deux types d'analyse ayant de nombreux liens de parenté mais légèrement différents: l'analyse en composantes principales et l'analyse factorielle proprement dite. Avant d'examiner les différences entre ces deux types d'analyse, il est pertinent de voir les points communs à la famille des analyses factorielles.

L'analyse factorielle cherche à **réduire un nombre important d'informations** (prenant la forme de valeurs sur des variables) **à quelques grandes dimensions**. Comme dans toute analyse statistique, on tente donc d'expliquer la plus forte proportion de la variance (de la covariance dans le cas de l'analyse factorielle) par un nombre aussi restreint que possible de variables (appelées ici composantes ou facteurs). On utilise le terme de variables latentes pour parler de ces variables qui existent sur le plan conceptuel seulement et qui ne sont pas mesurées.

L'analyse factorielle donne un sommaire des patrons de corrélations entre les variables. Elle tente de décomposer les patrons de corrélations pour les expliquer par un nombre restreint de dimensions. Elle est souvent utilisée comme méthode d'analyse exploratoire en vue de créer des échelles.

Exemples:

- De façon à mesurer la satisfaction des gens face à leur travail, j'ai d'abord déterminé que celle-ci portait sur trois grandes dimensions: la qualité des relations interpersonnelles, la nature même du travail et les aspects extrinsèques (salaire, horaire,...).

- Pour la dimension *nature du travail*, j'ai posé quatre (4) questions du type "Etes-vous très satisfait, assez satisfait, peu satisfait ou pas du tout satisfait a) des tâches qui vous sont assignées b) des défis que comporte votre travail c) de l'utilisation de vos compétences dans votre travail d) de votre capacité de vous réaliser dans votre travail.

- En agissant ainsi, je **suppose qu'une dimension générale de satisfaction face à la nature du travail existe** et que le positionnement des répondants face à cette dimension "explique", "prédit" leur positionnement sur chacune des "variables mesurées", soit les questions posées.

- **Si cette hypothèse est vraie,** les personnes auront tendance à répondre de la même manière aux quatre questions portant sur cette dimension. Leurs réponses à ces questions seront donc plus corrélées entre elles qu'avec les réponses aux autres variables liées à des dimensions/ facteurs différents.

- Cette "manière de voir" suppose aussi que l'on conçoit que les variables mesurées, les questions posées, constituent un échantillon de l'ensemble des variables/ questions aptes à mesurer le concept choisi.

Les équations

Pour ce qui est de l'analyse factorielle proprement dite, contrairement à l'entendement intuitif, il faut comprendre que ce sont les réponses aux variables mesurées qui dépendent des scores aux facteurs et non pas l'inverse. Ainsi, l'analyse postule que la réponse des individus à diverses questions portant, par exemple, sur leur satisfaction face à la nature de leur travail dépend de deux éléments: d'une part, leur satisfaction générale face à la nature de leur travail et d'autre part, un élément résiduel qui comprend l'erreur de mesure et un aspect unique propre à la satisfaction spécifique à l'élément sur lequel porte chaque question.

Les équations s'écrivent ainsi:

S'il y a un seul facteur et trois variables:

$x_1 = b_1F+u_1$
$x_2 = b_2F+u_2$
$x_3 = b_3F+u_3$

où: x_1 est la réponse donnée à la question 1
b_1 est la valeur du coefficient de régression et donne l'importance de l'influence de F – le facteur latent – sur x_1
F est la valeur théorique du facteur F
et u_1 est la valeur du résidu comprenant la part unique de variation propre à la variable x_1.

Similairement, s'il y avait n facteurs, l'équation pour une variable quelconque x_i s'écrirait:

$x_i= b_{i1}F_1 + b_{i2}F_2+...b_{in}F_n + u_i$

Ainsi, dans le cas par exemple d'un questionnaire de sondage ou d'un test, on peut dire que la réponse que donne un individu à une question posée est conçue comme une combinaison linéaire a) de la réponse théorique qu'il donnerait à une variable globale que l'on ne peut pas mesurer directement et b) d'autres facteurs reliés entre autres à la question spécifique qui est posée.

Il faut noter que ce qui précède vaut pour l'analyse factorielle et non pour l'analyse en composantes principales. Nous expliciterons plus loin la différence entre les deux mais il faut préciser ici que, dans l'analyse en composantes principales, les composantes – souvent appelés également facteurs – sont des combinaisons des variables mesurées. Celles-ci "causent" les composantes, contrairement à ce qui se passe en analyse factorielle.

Le type de questions auxquelles l'analyse factorielle permet de répondre (Tabachnik et Fidell, 2013: 615-616)

L'analyse factorielle permet de répondre à plusieurs questions. Il est important d'utiliser les ressources de l'analyse de façon à profiter de toutes ses possibilités. Voici le type de questions auxquelles l'analyse permet de répondre.

- *Combien de facteurs* sont nécessaires pour donner une représentation adéquate et parcimonieuse des données.

- Quelle *proportion de la variance* des données peut être expliquée par un certain nombre de dimensions (facteurs) majeures.

- Quelle est la *nature des facteurs*, comment on peut les interpréter.

- Jusqu'à quel point la solution factorielle proposée est conforme à la théorie que je voulais vérifier.

- La structure factorielle est-elle la même pour divers groupes?

- Quel score les répondants auraient vraisemblablement obtenu si on avait pu mesurer les facteurs directement.

L'analyse en composantes principales et l'analyse factorielle

La différence entre ces deux types d'analyse n'est pas toujours évidente, ceci d'autant plus que, suivant les "habitudes" disciplinaires et culturelles, certains ont tendance à utiliser systématiquement un ou l'autre de ces types d'analyse.

- L'analyse en composantes principales (ACP) cherche une solution **à l'ensemble de la variance des variables mesurées**. De plus, elle cherche une solution où les composantes sont orthogonales (c'est-à-dire indépendantes) entre elles[1]. Quelque soit la matrice de corrélations, il y a toujours une solution en ACP. L'ACP maximise la variance expliquée. Les composantes sont en quelque sorte une agrégation des variables corrélées (Tabachnik et Fidell, 2013, p. 615)

- L'analyse factorielle (A.F.) cherche une solution à **la covariance entre les variables mesurées**. Elle tente d'expliquer seulement la variance qui est commune à au moins deux variables et présume que chaque variable possède aussi une variance unique représentant son apport propre. Les divers

[1] Quoiqu'il soit possible de faire des rotations orthogonales ou obliques en ACP, cette utilisation ne respecte pas les bases mêmes de l'ACP, à savoir une solution unique et des composantes indépendantes entre elles qui expliquent chacune une proportion décroissante de la variance.

types d'extraction visent à maximiser une bonne reproduction de la matrice de corrélations originale en privilégiant certains aspects, différents pour les divers types.

Selon Tabachnik et Fidell (2013),

> "*Le choix entre l'analyse en composantes principales et l'analyse factorielle dépend de votre évaluation de l'adéquation entre les modèles, les données et le but de la recherche. Si vous êtes intéressés par une solution théorique non contaminée par la variance spécifique et la variance d'erreur et que vous avez élaboré votre recherche en vous basant sur des concepts précis qui devraient donner lieu à des scores spécifiques sur les variables observées, l'analyse factorielle est le choix approprié. Par contre, si vous voulez simplement un sommaire empirique de vos données, l'analyse en composantes principales est le choix approprié". P. 640.*

Notez que plusieurs auteurs, et en particulier les mathématiciens, semblent préférer l'analyse en composantes principales (voir par exemple Field, 2005, p. 630). Toutefois, en sciences sociales, c'est le modèle de l'analyse factorielle qui apparaît habituellement le plus approprié. Field a raison de noter toutefois que, dans la pratique, il est rare que les deux modèles donnent des structures factorielles différentes. Le texte de Lauri Tarkonen en annexe explique également l'historique de l'apparition des deux modèles et les raisons pour lesquels certains préfèrent un modèle plutôt qu'un autre.

L'analyse exploratoire et l'analyse confirmatoire

L'analyse habituellement effectuée par les logiciels courants est une analyse de type exploratoire. Elle ne permet pas de déterminer à l'avance quelles variables devraient être liées à quels facteurs. Lorsque la solution factorielle proposée par le logiciel (la solution statistique) confirme nos hypothèses de départ, c'est bon signe. Lorsque ce n'est pas le cas, ceci n'infirme pas nécessairement nos hypothèses, ceci parce qu'une multitude de solutions sont possibles pour chaque analyse et que le logiciel ne peut en proposer qu'une seule, celle qui est la plus appropriée statistiquement. Une autre solution, plus conforme à nos hypothèses, peut être presque aussi bonne que la solution proposée. On ne peut pas le vérifier.

L'analyse factorielle confirmatoire permet de déterminer non seulement le nombre de facteurs mais aussi l'appartenance de chaque variable à un ou plusieurs facteurs. Ce type d'analyse doit être utilisé avec précaution, lorsque l'on est vraiment à l'étape finale de la confirmation d'un modèle. Elle nécessite l'utilisation de logiciels permettant de faire des analyses par équations structurelles (EQS et LISREL, AMOS, un module de SPSS et un module de Stata 12). Tant LISREL qu'AMOS et STATA offrent une interface relativement conviviale. Attention, ceci peut vous amener à penser qu'il s'agit d'une analyse facile à effectuer où il suffit de donner les instructions au logiciel. Ce n'est pas le cas. Il faut vraiment savoir où on s'en va et avoir fait les analyses exploratoires avant d'utiliser ce type d'analyse. Il faut aussi bien comprendre le processus d'analyse.

2. Considérations théoriques et pratiques (*Tabachnik et Fidell, 2013: 616-620)*

- Pour qu'une variable soit intégrée dans l'analyse, sa distribution doit montrer une certaine variance: elle doit discriminer les positions des individus. Il faut donc examiner la distribution des variables avant de décider des variables à intégrer à l'analyse. La mesure n'est pas toujours conforme à nos attentes!

- Idéalement, on cherche une ***structure simple***, c'est-à-dire une solution où chaque variable est influencée fortement par ***un et un seul facteur***.

- Lorsqu'une variable est corrélée à plus d'un facteur, on dit que qu'il s'agit d'une ***variable complexe***; ce qui signifie que les réponses à cette variable s'interprètent selon deux dimensions ou même plus.

- La structure factorielle peut être différente pour différentes populations. Il faut faire attention à ne pas regrouper dans l'analyse des populations trop différentes.

- Il faut s'assurer que la solution factorielle soit *stable et généralisable*, ce qui nécessite de la valider sur un nombre suffisant de cas. **La convention veut qu'il y ait un strict minimum de 5 cas par variable.** Lorsque cette règle n'est pas respectée, plusieurs problèmes peuvent survenir dont celui de la "matrice malade" (ill-conditioned matrix). Un moins grand nombre de cas entraîne une moins grande confiance dans la possibilité de généraliser l'analyse: une deuxième analyse avec une population différente pourrait donner des regroupements différents. Il y a donc des problèmes de stabilité et de fiabilité de la solution factorielle.

- Les variables utilisées pour l'analyse devraient **se distribuer normalement**. Toutefois, lorsqu'on utilise l'analyse factorielle uniquement comme outil exploratoire, il est possible de "transgresser" cette règle et donc de faire une analyse factorielle sur des variables dont la distribution est très peu normale ou même avec des variables binaires de type (0,1). Il faut alors utiliser une procédure d'extraction (Moindres carrés non pondérés - ULS) qui tient compte du fait que la distribution des variables n'est pas normale.

- La relation entre les paires de variables est présumée **linéaire**.

- On devrait idéalement repérer et éliminer les cas ayant des patrons de réponses atypiques (**Outliers**).

- **La matrice de corrélation ne peut pas être singulière.** Ceci signifie que les variables ne peuvent pas être à ce point corrélées entre elles qu'une variable constitue une combinaison linéaire d'une ou de plusieurs autres variables; il y a alors "redite", c'est-à-dire que la même information est présente à deux reprises. Mathématiquement, les opérations sur les matrices nécessaires à l'estimation ne peuvent être effectués dans une telle situation.

- La matrice de corrélation doit contenir un patron, une solution factorielle. Certains ensembles de variables doivent être corrélés entre eux suffisamment pour qu'on puisse dire qu'ils dépendent d'un même facteur. La solution factorielle doit aussi expliquer une proportion suffisamment importante de la variance pour que la réduction à un nombre restreint de facteurs ne se fasse pas au prix d'une perte importante d'information. Tabachnik et Fidell (2013: 619) présentent divers moyens pour vérifier si la matrice de corrélation est appropriée pour l'analyse factorielle. Toutefois, les indices d'adéquation présentés plus loin sont habituellement suffisants pour évaluer si la matrice est appropriée.

- Toutes les variables doivent faire partie de la solution c'est-à-dire être corrélées substantiellement avec au moins une autre variable, sinon elles doivent être retirées de l'analyse puisqu'elles n'appartiennent pas à la solution factorielle. Le problème apparaît de façon évidente lorsque l'on examine les premiers tableaux de l'analyse factorielle puisque la qualité de la représentation sera trop basse et la variable ne sera pas corrélée à un facteur (voir plus loin comment repérer ces informations).

3. Les types d'extraction d'une solution factorielle.

L'extraction pour l'analyse en composantes principales (ACP)

Il s'agit d'une opération numérique et non d'une estimation. Ce type d'extraction produit nécessairement une solution et la solution produite est unique. Il s'agit d'un solution maximisant la variance expliquée par les facteurs.

L'extraction pour l'analyse factorielle (AF)

Il y a plusieurs méthodes d'extraction. Il faut souligner que lorsque la solution factorielle est stable, les diverses méthodes donnent des résultats similaires, sinon identiques.

Les méthodes les plus utilisées sont:

- Maximum de vraisemblance (ML pour Maximum Likelihood): maximise la probabilité que la solution factorielle reflète une distribution dans la population. Cette méthode produit aussi un test de χ^2 de maximum de vraisemblance qui indique si la solution factorielle retenue est généralisable à l'ensemble de la population. ***La probabilité de ce test doit être supérieure à 0,05, c'est-à-dire que l'on ne doit pas rejeter l'hypothèse nulle qui veut que le modèle soit compatible avec les données.***

Cette méthode est toutefois *sensible aux déviations à la normalité des distributions*: on rencontre fréquemment des problèmes lorsqu'on l'utilise avec des échelles ordinales de type Likert.

- Moindres carrés non pondérés (ULS ou Unweighted Least Square): minimise les résidus. Cette méthode est privilégiée lorsque les échelles de mesure sont ordinales ou que la distribution des variables n'est pas normale. Cette situation se présente fréquemment en sciences sociales,

particulièrement lorsque l'on mesure des attitudes. C'est pourquoi il s'agit de la méthode que l'on retrouve le plus souvent dans les articles, particulièrement en psychologie.

- **Alpha**: Cette méthode est très peu utilisée et peu connue. Elle s'avère pertinente lorsque le but de l'analyse est de créer des échelles puisqu'elle tente de maximiser l'homogénéité à l'intérieur de chaque facteur et donc, la fiabilité. Il est étonnant qu'elle soit si peu utilisée puisque l'objectif de beaucoup d'analyses factorielles est d'orienter la création d'échelles.

Il y a quelques autres types d'extraction qui ont tous certains avantages et certains inconvénients. Les méthodes présentées ici sont les plus fréquemment utilisées en sciences sociales.

4. Les types de rotation

La rotation est le processus mathématique qui permet de faciliter l'interprétation des facteurs en maximisant les saturations les plus fortes et en minimisant les plus faibles de sorte que chaque facteur apparaisse influencer un ensemble restreint et unique de variables. Ce processus est effectué par rotation, soit un repositionnement des axes.

Il existe deux principaux types de rotations:

La rotation orthogonale:

On utilise cette rotation lorsque l'on a de bonnes raisons de croire qu'il est possible d'extraire des facteurs qui soient indépendants les uns des autres. Une solution orthogonale **est toujours préférable parce qu'une telle solution indique que chaque facteur apporte une information unique, non partagée par un autre facteur. Toutefois, ce type de solution est rarement possible en sciences sociales puisque habituellement, il existe des liens conceptuels entre les facteurs**, ce qui entraîne que les facteurs sont corrélés entre eux. Il existe plusieurs méthodes pour produire une rotation orthogonale; la plus fréquemment utilisée et probablement la plus stable est **VARIMAX.**

La rotation oblique:

La rotation oblique permet qu'il y ait corrélation entre les facteurs. Comme elle correspond habituellement mieux à la "réalité", soit une interrelation entre les diverses dimensions, elle est fréquemment utilisée en sciences sociales. La méthode utilisée est **OBLIMIN.** On peut maintenant fixer la corrélation maximale entre les facteurs que l'on est prêt à accepter en donnant une valeur à delta (de -4, pratiquement orthogonale, à 0, par défaut dans oblimin, où les facteurs peuvent être assez corrélés, à 1, où les facteurs peuvent être très corrélés).

5. Les étapes de l'analyse factorielle de type exploratoire

a) Sélectionner l'ensemble des variables qui seront analysées conjointement.

b) Idéalement, examiner cet ensemble de façon conceptuelle et déterminer la solution qui apparaîtrait plausible quant au nombre de facteurs et au regroupement des variables.

c) Effectuer une analyse factorielle avec une rotation oblique (oblimin). Il s'agit ici de voir s'il est possible d'obtenir une structure simple ou chaque variable est liée à un et un seul facteur. La rotation oblique permet de savoir si les facteurs sont corrélés entre eux. Si les corrélations sont faibles, moins de 0,32 selon Tabachnick et Fidell (2013:651), on peut passer à une rotation orthogonale.

Deux critères permettent de déterminer le nombre de facteurs à retenir. Le ***critère de Kaiser*** **(par défaut dans SPSS)** garde les facteurs dont la valeur propre est plus grande que 1, soit la variance d'une variable standardisée. L'idée est qu'un facteur doit avoir au moins la variance d'une variable**.** **Le *test du coude de Cattell ou test de l'éboulis*** peut également être utilisé. Il s'agit d'un test fait de façon visuelle à partir d'une représentation graphique des valeurs propres. On considère que le nombre de facteurs idéal est déterminé par la présence d'un "coude" au delà duquel les facteurs supplémentaires apportent peu à l'explication. Le coude peut aussi être compris comme l'endroit dans la courbe où la droit entre les points change de direction. Lorsque le test de l'éboulis indique un nombre de facteurs différent de ce qu'indique le critère de Kaiser, il est possible de vérifier si cette solution est préférable en spécifiant -- dans SPSS – le nombre de facteurs voulu. Le test de l'éboulis, aidé par des repères visuels, est également utilisé dans d'autres types d'analyse comme l'analyse des correspondances ou l'analyse de classification. Selon Tabachnick et Fidell (2013:649), le nombre de facteurs devrait se situer entre le nombre de variables divisé par 3 et le nombre de variables divisé par 5.

- S'il y a des problèmes de convergence dans l'analyse factorielle, on peut tenter une analyse en composantes principales pour voir si on obtiendra une solution. Il faut également examiner la matrice de corrélation pour détecter des corrélations très élevées de même que l'information sur la qualité de la représentation. Lorsque celle-ci est très très élevée, près de 1, il y a sans contredit un problème de multicollinéarité.

d) Examiner la première analyse pour évaluer les éléments suivants:

- Comparer la solution proposée par l'analyse avec l'hypothèse de regroupement faite au départ. Examiner sa pertinence théorique.

- Examiner les divers indices de pertinence de la solution factorielle (voir outils diagnostiques plus loin), particulièrement l'indice KMO.

- Pour chacune des variables, décider du maintien dans les analyses subséquentes à partir des critères suivants, examinés simultanément:

- *Vérifier si la qualité de la représentation* (statistiques initiales dans l'analyse factorielle) est suffisamment importante (>0,20) pour le maintien dans l'analyse. En bas de 0,20, la variable a une corrélation très faible avec les autres variables.

- *Vérifier si chaque variable est corrélée à au moins un facteur* avec une saturation factorielle plus grande que 0,30. Une variable qui a une faible qualité de représentation et en plus n'est pas suffisamment liée à au moins un facteur devrait être retirée presque automatiquement. Elle pourra être utilisée individuellement – hors échelle – dans les analyses subséquentes.

- *Examiner les variables qui ont une saturation sur plus d'un facteur* (variables complexes). Il faut comprendre pourquoi c'est le cas. Si la saturation est nettement plus forte sur un facteur que sur un autre, on peut penser que les analyses de fiabilité subséquentes permettront de maintenir la variable sur l'échelle correspondant au facteur principal. Si les saturations sont à peu près égales, il faut habituellement retirer la variable puisque cela signifie que sa signification est ambigüe.

- *Examiner parallèlement la pertinence théorique et conceptuelle* de maintenir ou de retirer une variable plutôt qu'une autre.

e) Refaire l'analyse de façon itérative jusqu'à arriver à une *solution simple* satisfaisante. Attention, enlever les variables problématiques une par une et non toutes ensemble, d'un seul coup. (J'appelle ça le "massacre à la scie à chaîne"). Il est possible que le simple fait de retirer une variable règle d'autres problèmes et permette de garder plus de variables dans l'analyse.

Une *solution simple* est un tableau de saturations factorielles indiquant que chaque variable est liée à un facteur avec une saturation plus grande que 0,30 et qu'elle est liée à un seul facteur. Il s'agit d'une indication de la validité convergente et discriminante de l'ensemble des mesures.

Le test d'une bonne analyse factorielle réside, en fin de compte, dans la signification des résultats. C'est le chercheur qui "décode" la signification conceptuelle de chaque facteur. Il faut nommer chaque facteur en fonction du concept auquel ils réfèrent tel qu'on peut le déduire des variables qui lui sont liées. *Évitez de numéroter vos facteurs.* Les numéros de facteur ne veulent rien dire et rendent votre solution incompréhensible pour les lecteurs. À l'opposé, il faut se garder de donner aux facteurs des noms qui font sens et qui impressionnent mais qui ne reflètent pas ce qui a été mesuré.

Deux problèmes se posent:
- Le critère de la justesse de l'analyse est en partie subjectif: Est-ce que le regroupement fait sens?
- *Il y a une infinité de solutions possibles après rotation.* il n'y a donc pas une seule "bonne" solution. Il est donc difficile de décréter que la solution présentée est "la solution". Il faut la présenter comme une solution plausible qui n'est pas contredite par les données.

6. Outils de diagnostic de la solution factorielle

Si le principal critère d'un bonne solution factorielle demeure sa justesse sur le plan théorique, sur le plan du sens, ***il demeure que plusieurs outils statistiques nous guident dans la recherche de la solution la plus appropriée.*** Voici une brève présentation des principaux outils diagnostiques utilisés. L'exemple présenté permettra d'en comprendre l'utilisation de façon plus poussée.

Adéquation de la solution globale

a) Le déterminant de la matrice:

Un déterminant égal a zéro signifierait qu'au moins une variable est une combinaison linéaire parfaite d'une ou de plusieurs autres variables. Il y a donc une variable qui ne rajoute aucune information nouvelle au-delà de celle fournie par les autres variables. Dans ce cas l'analyse ne peut pas procéder pour des raisons mathématiques (Il est impossible d'inverser la matrice). Si le déterminant est trop élevé, il s'agit du problème inverse soit que certaines variables ne sont pas assez corrélées à d'autres pour qu'une solution factorielle soit appropriée. *Nous cherchons donc à avoir un* ***déterminant très petit****, ce qui constitue un bon indice de l'existence de patrons de corrélations entre les variables,* ***mais non strictement égal à zéro (ce qui signifierait que la matrice est singulière). Notez que dans ce cas, l'analyse ne pourrait procéder dû à l'impossibilité d'inverser les matrices****.*

On obtient le déterminant
- en cochant *Déterminant* dans la fenêtre DESCRIPTIVES (SPSS Windows).

b) La mesure de Kaiser-Meyer-Olkin (KMO)

Plus communément appelé le KMO, la mesure de Kaiser-Meyer-Olkin est un indice d'adéquation de la solution factorielle. Il indique jusqu'à quel point l'ensemble de variables retenu est un ensemble cohérent et permet de constituer une ou des mesures adéquates de concepts. Un KMO élevé indique qu'il existe une solution factorielle statistiquement acceptable qui représente les relations entre les variables.

Une valeur de KMO de moins de .5 est inacceptable
.5 est misérable
.6 est médiocre
.7 est moyenne
.8 est méritoire
.9 est merveilleuse (ref: SPSS professional statistics)

Le KMO reflète le rapport entre d'une part les corrélations entre les variables et d'autre part, les corrélations partielles, celles-ci reflétant l'unicité de l'apport de chaque variable.

On obtient le KMO
- en cochant *Indice KMO et test de sphéricité de Bartlett* dans la fenêtre DESCRIPTIVES (SPSS Windows).

c) Le test de sphéricité de Bartlett:

Ce test vérifie l'hypothèse nulle selon laquelle toutes les corrélations seraient égales à zéro. On doit donc tenter de rejeter l'hypothèse nulle et la probabilité d'obtenir la valeur du test doit donc être plus petite que 0,05. Toutefois le test est très sensible au nombre de cas; il est presque toujours significatif lorsque le nombre de cas est grand. Ses résultats sont donc intéressants presque uniquement lorsqu'il y a moins de 5 cas par variable.

On obtient le test de sphéricité automatiquement
- en cochant *Indice KMO et test de sphéricité de Bartlett* dans la fenêtre DESCRIPTIVES (SPSS Windows).

d) Le test du coude de Cattell (ou test de l'éboulis)

Le graphique des valeurs propres donne une représentation graphique des informations sur les valeurs propres de chaque facteur présentées dans le tableau des statistiques initiales. Dans cette représentation, il faut rechercher le point (parfois les points) de cassure qui représente le nombre de facteurs au-delà duquel l'information ajoutée est insignifiante et peu pertinente. *Plus la courbe est accentuée, plus il apparaît qu'un petit nombre de facteurs explique la majeure partie de la variance. A partir du moment où la courbe devient presque une ligne droite horizontale, il apparaît que les facteurs subséquents apportent peu de nouvelles informations.*

Note: Les valeurs propres représentent la variance expliquée par chaque facteur. Elles sont constituées de la somme des poids factoriels au carré de toutes les variables pour un facteur déterminé.

On obtient cette représentation
- en cochant *Diagramme des valeurs propres* dans la boîte "afficher" de la fenêtre EXTRACTION (SPSS Windows).

e) La matrice reproduite et les résidus

L'analyse en composantes principales, tout comme l'analyse factorielle, constitue une décomposition de la matrice des corrélations entre les variables.

Ainsi, si l'on effectue une analyse en composantes principales et que l'on demande autant de facteurs qu'il y a de variables dans l'analyse, la matrice de corrélation reproduite sera identique à la matrice de corrélation initiale.

Ce n'est pas le cas pour l'analyse factorielle puisque celle-ci tente d'expliquer non pas la variance totale mais uniquement la covariance entre les variables. Donc, même avec autant de facteurs que de variables, **la matrice qui est créée lorsque l'on tente l'opération inverse,** c'est-à-dire de reproduire les corrélations d'origine à partir des informations extraites des facteurs suite à l'analyse, **ne reproduira pas les corrélations originales à la perfection, il restera des résidus. Il reste d'autant plus de résidus que l'on garde seulement une partie des facteurs.**

Plus la solution factorielle est bonne, plus la matrice "reproduite" s'approche de la matrice de corrélation originale et moins les résidus sont importants. L'indication d'une proportion faible de résidus standardisés plus grands que 0,05 est un indice d'une bonne solution factorielle qui représente bien toutes les corrélations entre les variables. Les résidus plus grands que 0,05 devraient être examinés pour voir s'il n'y a pas des cas "aberrants" et de tenter de comprendre pourquoi. On pourrait avoir à éliminer une variable de l'analyse.

On obtient la matrice reproduite et la matrice des résidus
- en cochant *Reconstituée* dans la boîte "matrice de corrélation" de la fenêtre DESCRIPTIVES (SPSS Windows).

f) La structure obtenue

La structure obtenue, c'est-à-dire le tableau des corrélations entre les variables et les facteurs (*matrice des facteurs après rotation* en rotation orthogonale et *matrice des types* en rotation oblique), doit être ***simple***, **ce qui veut dire que chaque variable doit avoir une corrélation plus grande que 0,3 avec au moins un facteur et avec un seul facteur.**

Ces matrices sont imprimées automatiquement
- en cochant *Structure après rotation* dans la fenêtre ROTATION (SPSS Windows). Pour avoir ces mêmes tableaux avant rotation, on coche *Structure factorielle sans rotation* dans la fenêtre EXTRACTION.

Pertinence de garder une variable dans la solution

a) Les statistiques descriptives des variables

Les variables, par définition, doivent montrer une certaine variation du positionnement des individus quant à ce qui est mesuré. En ce sens, un écart-type important et une moyenne qui se rapproche du milieu de l'échelle de mesure (exemple: moyenne de 2.5 pour une échelle à 4 catégories) sont de bons indices que la variable apporte une information susceptible d'aider à différencier les individus.

Ces statistiques sont produites
- en cliquant *Caractéristiques univariées* dans la fenêtre DESCRIPTIVES (SPSS Windows).

b) La qualité de la représentation de chaque variable avec la solution factorielle initiale

On doit examiner les statistiques de qualité pour l'analyse factorielle(et non l'analyse en composantes principales puisque cette qualité est de 1,0 par définition) et ceci avant l'extraction d'un nombre restreint de facteurs. Cette information représente l'appartenance de chaque variable à la covariance de l'ensemble des variables. **C'est la proportion de la variance de chaque variable qui peut être expliquée par l'ensemble des autres variables.** On considère que la qualité doit être d'au moins 0,20 pour justifier le maintien de la variable dans l'analyse. En analyse factorielle, on peut toutefois tenir également compte de la qualité de la représentation après extraction. Si celle-ci est suffisamment élevée et que les autres indices sont bons, on peut penser à garder la variable dans l'analyse même si sa qualité de représentation avant extraction est plus faible que 0,20.

Par ailleurs, il pourrait également arriver qu'une variable ait une bonne qualité de représentation avec la solution initiale mais non avec la solution après extraction d'un nombre restreint de facteurs. Ceci se refléterait probablement par le fait que cette variable ne se regrouperait pas avec les autres dans la solution factorielle (elle ne serait probablement pas corrélée suffisamment avec un facteur). L'inverse peut aussi se produire, auquel cas on aura tendance à maintenir la variable dans l'analyse puisqu'elle appartient à la solution après extraction.

NOTE: La somme des poids factoriels au carré pour une variable donnée égale la qualité de représentation de cette variable.

On obtient ces informations
- en cochant *Structure initiale* dans la fenêtre DESCRIPTIVES (SPSS Windows).

c) La simplicité ou la complexité de chaque variable dans la solution factorielle finale

Une variable est dite complexe lorsqu'elle est corrélée substantiellement (saturation factorielle plus grande que 0,30 à plus d'un facteur. On peut dire que les réponses à cette variable reflètent plus d'un concept. Ainsi, il pourrait arriver que la satisfaction face aux avantages sociaux soit corrélée à deux facteurs, un portant sur la rémunération (associé au salaire) et l'autre portant sur la sécurité d'emploi (puisque les avantages sociaux sont meilleurs lorsqu'il y a sécurité d'emploi). Cela entraîne un problème lorsque l'on veut créer des échelles : Avec quelle échelle devrait-on regrouper cette variable, celle portant sur la rémunération ou celle portant sur la sécurité?

Il y a plusieurs manières de traiter ce problème dépendant de l'importance théorique de la variable et du choix quant au nombre de facteurs. Dans le cas où d'autres variables amènent une information similaire, on peut retirer la variable considérée comme complexe. Il arrive qu'on la maintienne dans l'analyse; lorsque l'on veut créer les échelles, on décide de son appartenance à une échelle plutôt qu'une autre à partir de la saturation factorielle la plus élevée, des informations fournies par les analyses de fidélité et à partir de considérations théoriques.

Ces informations sont tirées des matrices factorielles (Matrice des facteurs ou matrice des types) que l'on obtient
- en cochant *Structure après rotation* dans la fenêtre ROTATION (SPSS Windows).

d) Un cas spécial: le cas Heywood

On parle de cas Heywood lorsque, dû aux relations trop fortes entre certaines variables, un problème se produit dans le processus d'estimation et la saturation factorielle devient plus grande que 1,0. Comme il s'agit d'une corrélation, ceci devrait être impossible. Cette situation se produit plus fréquemment avec la rotation oblique et avec des modes d'extraction comme le maximum de vraisemblance. On repère facilement la présence d'un cas Heywood

- lorsque la qualité d'une variable est notée comme étant très élevée (0,9996 ou plus).
- lorsque la saturation factorielle d'une variable est plus grande que 1,0. (SPSS émet un avertissement disant que la qualité d'une ou de plusieurs variables est plus grande que 1,0).

Le cas Heywood est dû au fait que la variable est trop fortement corrélée à une ou plusieurs autres variables. Dans ce cas, il faut décider soit de retirer la variable des analyses subséquentes soit de retirer une autre variable avec laquelle elle est fortement corrélée.

7. Commandes pour l'analyse factorielle avec SPSS

En résumé, une commande d'analyse factorielle dans SPSS qui donnerait l'ensemble des informations nécessaires présentées plus haut aurait la forme suivante:

COMMANDE, SOUS-COMMANDE	SIGNIFICATION
FACTOR /VAR=NATURE3 TO RECON8 CARRIER1 STABIL2 PERFECT9 SECUR13 SALAIR10 TO HORAIR12 ** Dans Windows, sélectionner les variables*. Il y a également maintenant la sous-procédure /ANALYSIS. Ceci permet, lorsque l'on édite la syntaxe, de modifier uniquement l'ensemble de variables soumis à l'analyse sans toucher à l'ensemble initial.	FACTOR: demande une analyse factorielle /VAR= liste des variables qui seront analysées
/PRINT DEFAULT UNIVARIATE CORRELATION REPR DET KMO **Dans Windows, choisir Structure initiale, Caractéristiques univariées, Coefficients, Reconstituée, Déterminant et Indice KMO et test de Bartlett dans CARACTÉRISTIQUES*	/PRINT indique les informations qui devront être imprimées, dans ce cas-ci, les statistiques par défaut, les statistiques univariées pour chaque variable, la matrice de corrélation originale, la matrice reproduite, le déterminant, le KMO et le test de sphéricité.
/CRITERIA FACTORS (4) ** Dans Windows, cette option est disponible dans la fenêtre EXTRACTION*	/CRITERIA permet de spécifier, *lorsque nécessaire,* des critères tels le nombre de facteurs et le nombre d'itérations; dans ce cas-ci, on demande 4 facteurs.
/PLOT EIGEN ROTATION ** Dans Windows, on demande le Graphique des valeurs propres dans la fenêtre EXTRACTION et Carte factorielle dans la fenêtre ROTATION. Par défaut , cette dernière indication donne un graphique en 3 dimensions des 3 premiers facteurs, le cas échéant.*	/PLOT permet de demander des graphiques des valeurs propres ou des facteurs; dans ce cas-ci on demande le graphique des valeurs propres et celui des deux premiers facteurs

/FORMAT SORT BLANK (.3) ** Dans Windows, la fenêtre OPTIONS donne l'Affichage des projections. On clique* ***Classement des variables par taille*** *et Supprimer les valeurs absolues inférieures à __. On change la valeur par défaut (.10) à .30.*	/FORMAT contrôle l'apparence; dans ce cas-ci SORT permet que les variables apparaissent dans les matrices factorielles en fonction de leur importance et selon les facteurs et BLANK (.3) permet que les saturations factorielles inférieures à .3 n'apparaissent pas dans l'impression ce qui facilite la lecture.
/EXTRACTION ULS /ROTATION OBLIMIN. **Dans SPSS Windows, il faut, pour chaque extraction ou rotation, refaire la commande d'analyse factorielle au complet (i.e. choisir le mode d'extraction approprié dans EXTRACTION et la rotation désirée dans ROTATION). Attention, c'est dans la fenêtre EXTRACTION que l'on demande le Graphique des valeurs propres ainsi que la solution factorielle sans rotation et que l'on définit les critères (nombre de facteurs et/ou d'itérations) le cas échéant. Dans ROTATION, on demande la Structure après rotation et les cartes factorielles.*	EXTRACTION spécifie le type d'extraction et ROTATION le type de rotation. Dans ce cas-ci, on demande une analyse avec extraction ULS (moindres carrés non pondérés) comprenant une rotation une rotation oblique.

En résumé, voici les commandes pour l'analyse factorielle avec Spss -Windows:
→ Allez dans ANALYSE,
→ choisir RÉDUCTION DES DIMENSIONS - ANALYSE FACTORIELLE
Dans le tableau principal de l'analyse factorielle

a) Choisir les VARIABLES que l'on veut analyser

b) Dans CARACTÉRISTIQUES : - STATISTIQUES→ CARACTÉRISTIQUES UNIVARIÉES
→ STRUCTURE INITIALE
- MATRICE DE CORRÉLATIONS
→ COEFFICIENTS
→ DÉTERMINANT
→ INDICE KMO ET TEST DE BARTLETT
→ MATRICE DES CORRÉLATION RECONSTITUÉE

c) Dans EXTRACTION :
- MÉTHODE →COMPOSANTES PRINCIPALES (PC)
→MOINDRES CARRÉS NON PONDÉRÉS (ULS)
→ALPHA-MAXIMISATION
→...
- EXTRAIRE →VALEURS PROPRES SUPÉRIEURES À 1.0
→NOMBRE DE FACTEURS =
→MAXIMUM DES ITÉRATIONS POUR CONVERGER (pour augmenter le nombre d'itérations nécessaires à la convergence, au besoin).
- AFFICHER →STRUCTURE FACTORIELLE SANS ROTATION
→GRAPHIQUE DES VALEURS PROPRES

d) Dans ROTATION:
- MÉTHODE →VARIMAX
→OBLIMIN DIRECTE
→AUTRES
- AFFICHER →STRUCTURE APRÈS ROTATION
→CARTE(S) FACTORIELLE(S) (par défaut: 3-D pour 3 premiers facteurs)

e) Dans OPTIONS:
- VALEURS MANQUANTES
→ EXCLURE TOUTE OBSERVATION INCOMPLÈTE OU EXCLURE SEULEMENT LES COMPOSANTES NON VALIDES OU *REMPLACER PAR LA MOYENNE (PRÉFÉRÉ)*
- AFFICHAGE DES PROJECTIONS
→CLASSEMENT DES VARIABLES PAR TAILLE (TR. IMPORTANT)
→SUPPRIMER LES VALEURS ABSOLUES INFÉRIEURES À (.30) ON PEUT LE METTRE À .25 POUR VOIR SI CERTAINES VARS. ONT UNE SATURATION PROCHE DE ,30)

8. Qu'est ce que l'analyse de fidélité (fiabilité)?

L'analyse de la fiabilité est une suite "normale" à l'analyse factorielle. En effet, suite à l'analyse factorielle, on aura identifier un certain nombre de facteurs. Quoique l'on puisse utiliser les scores factoriels produits par l'analyse factorielle dans certaines conditions comme mesures des facteurs, la pratique est plutôt de créer des échelles à partir de l'information fournie par l'analyse factorielle. Ces nouvelles variables pourront être utilisées dans les analyses subséquentes, entre autres dans les régressions. On voudra donc créer de nouvelles variables en combinant les réponses aux variables qui sont associées à *chaque* facteur. Pour cela, il faut vérifier si les mesures composites que nous voulons créer seraient fiables. Habituellement, cette vérification donne des résultats qui confirment ceux de l'analyse factorielle. Mais, comme l'analyse factorielle prend en compte uniquement la variance commune aux variables et que les nouvelles mesures composites seront créées à partir de l'ensemble de la variance, il peut arriver qu'il y ait contradiction entre les deux types d'analyse. Dans ce dernier cas, il faudra souvent retourner à l'analyse factorielle pour chercher une nouvelle solution. L'avantage de l'analyse de fidélité est de donner une mesure de la fiabilité des échelles que nous utiliserons dans les analyses subséquentes.

Blalock (1968):

"Les sociologues théoriciens utilisent souvent des concepts qui sont formulés à un assez haut niveau d'abstraction. Ce sont des concepts relativement différents des variables utilisées qui sont le lot des sociologues empiriques... Le problème du lien entre la théorie et la recherche peut donc être vu comme une question d'erreur de mesure".

La mesure peut être vue comme le "processus permettant de lier les concepts abstraits aux indicateurs empiriques" (Carmines et Zeller, 1979).

Deux concepts clés: La fidélité (ou fiabilité) et la validité

Fidélité: *Consistance dans la mesure:* Jusqu'à quel point plusieurs mesures prises avec le même instrument, par exemple un questionnaire, donneront les mêmes résultats dans les mêmes circonstances.

Exemple: Je fais passer un questionnaire portant sur l'idéologie deux fois aux mêmes personnes à deux mois d'intervalle et j'obtiens des résultats différents entre les deux passations. Est-ce que l'idéologie d'une personne peut changer si vite ou si c'est l'instrument qui n'est pas fiable?

La fidélité – ou fiabilité – demeure sur le plan empirique: elle indique si, en soi, l'instrument est un bon instrument.

Validité: *Jusqu'à quel point l'instrument mesure ce qu'il est supposé mesurer.*

Exemple: Si j'utilise une série de questions visant à mesurer les préférences idéologiques et que je me rends compte que j'ai en fait mesuré l'identification à un parti politique, ma mesure est une mesure non valide de l'idéologie.

La validité concerne la relation entre la théorie et les concepts qui lui sont reliés d'une part et la mesure d'autre part: elle est concernée par l'adéquation de la traduction du concept en mesure.

Il y a plusieurs types de validité (Durand et Blais, 2009: 244-246):

validité de contenu: relation entre le - les concepts à mesurer et l'instrument utilisé. Un instrument de mesure de l'aliénation mesurera le sentiment d'absence de pouvoir, d'absence de normes, d'isolation sociale, etc. Ceci exige que le domaine et les concepts soient bien définis au départ.

validité convergente et discriminante: Jusqu'à quel point chaque indicateur constitue une mesure d'un et d'un seul concept.

validité de construit: peut être validée par la relation entre l'instrument et d'autres instruments supposés mesurer des concepts reliés.

validité reliée au critère*:** relation entre l'instrument et ce à quoi il devrait théoriquement être relié. On parle de *validité **prédictive quand le critère est mesuré après et de *validité **concurrente*** quand le critère est mesuré en même temps.

La fidélité (ou fiabilité):

→ théorie classique des tests:

→ le score observé (qui peut-être par exemple la réponse d'une personne à une question) est une combinaison linéaire d'une partie représentant le score vrai et d'une partie d'erreur aléatoire.

$x_i=t_i+e_i$
où "x_i" représente la réponse de l'individu à une question
"t_i" représente le score vrai
et "e_i" représente l'erreur aléatoire soit l'écart entre la réponse de l'individu et le score vrai.

→ la corrélation entre deux scores observés constitue un estimé de la fidélité des mesures. → Dès qu'on a plus d'une mesure d'un même concept, on peut estimer la fidélité ρ; on peut concevoir l'erreur moyenne de mesure comme la réciproque de la fidélité, "$1-\rho$ ". Dans la théorie classique des tests, ceci est vrai si les scores

vrais et la variance d'erreur sont identiques (i.e. les mesures sont parallèles).

Diverses manières de mesurer la fidélité:

fidélité test-retest: corrélation entre la mesure prise à un temps 1 et la mesure prise à un temps 2: fidélité dans le temps.

fidélité "split-half" entre deux sous-ensembles: Jusqu'à quel point deux sous-ensembles des items constituent deux mesures fidèles du même concept.

fidélité entre différentes formes: Jusqu'à quel point deux ensembles différents d'items peuvent mesurer le même concept.

consistance interne: Jusqu'à quel point chacun des items constitue une mesure équivalente d'un même concept.

On peut mesurer la consistance interne en utilisant le

→ *Alpha de Cronbach:*

$$\alpha = \frac{k * \bar{cov}/\bar{var}}{1 + (k-1)\bar{cov}/\bar{var}}$$

où "k" est le nombre d'items
$\bar{cov}$ est la covariance moyenne entre les items
et $\bar{var}$ est la variance moyenne des items

Si les items sont standardisés de façon à avoir la même variance, la formule se modifie comme suit:

$$\alpha = \frac{k * \bar{\rho}}{[1 + \bar{\rho}(k-1)]}$$

où "k" représente le nombre d'items dans l'échelle
et "$\bar{\rho}$" est la corrélation moyenne

Ce coefficient Alpha peut être considéré comme la moyenne des coefficients alpha que l'on obtiendrait pour toutes les combinaisons possibles de deux sous-ensembles des items mesurant un même concept. Il peut aussi être vu comme l'estimé de la corrélation que l'on obtiendrait entre un test et une forme alternative du même test comprenant le même nombre d'items.
Le coefficient alpha est la borne inférieure de la fidélité réelle i.e la fidélité réelle ne peut pas

être inférieure à la valeur du alpha et elle est égale à cette valeur lorsque les items sont parallèles i.e. les scores vrais ont la même moyenne et la variance d'erreur est la même.

Remarque: La valeur de alpha augmente avec le nombre d'items, mais ce à la condition que la corrélation moyenne inter-item ne soit pas diminuée avec l'ajout de nouveaux items (et donc toutes choses égales par ailleurs). L'amélioration du alpha devient marginale au-delà d'un certain nombre d'items (environ 6-7).

9. L'analyse de fidélité : aspects pratiques et outils diagnostiques

Plusieurs outils sont disponibles pour évaluer la fidélité d'un ensemble de variables. La procédure RELIABILITY de SPSS permet d'examiner les informations pertinentes. On l'obtient en cliquant "analyse de fiabilité" dans le menu "Echelle".

Matrice de corrélation: Tout comme avec l'analyse factorielle, cette matrice permet de voir jusqu'à quel point les items sont corrélés entre eux et quels items sont plus fortement corrélés. S'il s'avérait que deux concepts sont mesurés plutôt qu'un seul, les corrélations pourraient nous permettre de repérer cette possibilité. ***Notez qu'il ne doit pas y avoir de corrélation faible et surtout pas de corrélations négatives dans cette matrice.*** Des corrélations négaitves indiqueraient que certaines variables devraient être inversées parce que leur relation avec les autres variables va en sens inverse.

→ Dans SPSS Windows, on coche *Corrélations* dans la boîte "Récapitulatifs" des statistiques.

Statistiques univariées:
Pour chacun des items, on peut obtenir la moyenne et l'écart type, ce qui permet de voir si les statistiques descriptives sont similaires pour les divers items.

→ Dans SPSS Windows, on coche *Item* dans "Caractéristiques pour" des statistiques.

Les statistiques d'échelle: indiquent quelle serait la moyenne, la variance et l'écart-type de l'échelle si on additionnait les réponses à chacun des items. Elles donnent une idée des propriétés futures de l'échelle à créer →Par exemple, la variance sera-t-elle suffisante?

→ Dans SPSS Windows, on coche *échelle* dans "Caractéristiques pour" des statistiques.

Le sommaire des statistiques d'items: les moyennes: Donnent les indications sur les différences de moyennes entre les items i.e minimum, maximum, moyenne, ratio maximum/minimum. Si ces différences sont trop importantes, on pourrait penser que chaque item ne mesure pas le concept de façon équivalente. On s'attend à un ratio maximum/minimum plus petit que 2.
→ Dans SPSS Windows, on coche *Moyennes* dans "Récapitulatifs" des statistiques.

Le sommaire des statistiques d'items: les variances: Donnent les indications sur les différences de variances entre les items. Même interprétation que pour les moyennes. ON s'attend également à un ratio maximum/minimum plus petit que 2.

→ Dans SPSS Windows, on coche *Variances* dans "Récapitulatifs" des statistiques.

Les statistiques d'items: les corrélations inter-items: Donnent les indications **très importantes** sur les différences de corrélations entre les items. Des corrélations très faibles ou négatives devraient être repérées dans la matrice de corrélation. Si cette situation existe, soit plus d'un concept est mesuré, soit certains items mesurent mal le concept ou l'échelle aurait due être inversée pour cet item (Par exemple, dans le cas où des items sont formulés négativement et d'autres positivement).

→ Dans SPSS Windows, on coche *Corrélations* dans "Cohérence inter-éléments" des statistiques

Les statistiques de la relation entre chaque item et l'échelle: "échelle sans l'élément":

Moyenne, variance qui résulterait si on enlevait un item: La moyenne et la variance de l'échelle ne devraient pas être fortement modifiés par le retrait d'un item sinon, ceci signifie que l'item en question ne contribue pas de la même manière à la mesure du concept que les autres items.

Corrélation corrigée: Il s'agit de la corrélation entre l'item et les autres items de l'échelle (i.e l'échelle moins l'item en question). Cette corrélation devrait être minimalement de 0,30 (comme dans le cas de la corrélation entre chaque item et le facteur dans l'analyse factorielle).

Corrélation multiple au carré: Il s'agit de la variance de l'item expliquée par les autres items. Plus elle est élevée, plus l'item est une mesure commune du concept. Il s'agit de l'équivalent de la qualité de la représentation dans les statistiques initiales en analyse factorielle.

Alpha si l'item était enlevé: Il est très important de regarder cette information. Si un item n'appartient pas à l'échelle, alors la fidélité telle que mesurée par le alpha **serait supérieure si l'item était enlevé**. Donc, ***lorsque l'alpha si l'item était enlevé est supérieur au alpha standardisé***, ceci signifie que l'item en question est une mesure qui détériore la fiabilité de l'échelle. Si la différence entre les deux alphas est appréciable, l'item doit automatiquement être retiré pour la suite des analyses.

Toutes les statistiques précédentes sont obtenues
Dans SPSS Windows, on coche *Échelle sans l'élément* dans "Caractéristiques pour" des statistiques
ALPHA ET ALPHA STANDARDISÉ: Ces deux mesures sont similaires lorsque les moyennes et les variances des items diffèrent peu. Sinon, il faut plutôt se fier au alpha standardisé. L'évaluation d'un bon alpha est similaire à celle du KMO présenté en analyse factorielle: On

recherche une valeur supérieure à .70, une valeur supérieure à .90 étant magnifique.

10. L'analyse de fidélité avec SPSS

En résumé, voici les commandes pour l'analyse de fidélité avec Spss -Windows:
→ Allez dans STATISTICS,
→choisir ECHELLE - ANALYSE DE FIABILITÉ

Dans le tableau principal de l'analyse de fidélité
a) Choisir les VARIABLES (items) que l'on veut analyser pour UNE échelle donnée et donner un nom approprié à l'échelle.

b) ***Entrer une étiquette d'échelle*** : CECI VOUS PERMETTRA DE VOUS REPÉRER PLUS TARD.

c) Dans MODÈLE: - choisir le type d'analyse → ALPHA DE CRONBACH
- autres choix (split-half, Guttman, parallèle, parallèle stricte)

d) Dans STATISTICS :
- CARACTÉRISTIQUES POUR
  → ITEM
  → ECHELLE
  → ECHELLE SANS L'ÉLÉMENT
- RÉCAPITULATIFS
  → MOYENNES
  → VARIANCES
  → CORRELATIONS
  → (COVARIANCES)
- COHÉRENCE INTER-ÉLÉMENTS
  → CORRELATIONS
  → (COVARIANCES)
- TABLEAU ANOVA
  → AUCUNE
  → (F TEST - teste qu'il n'y a pas de différence significative entre les différentes mesures de l'échelle)

En résumé, la syntaxe:

```
RELIABILITY
 /VARIABLES=q65c q65d q65e q65f q65g q65h
 /FORMAT=LABELS   /SCALE(ALPHA)=ALL
 /MODEL=ALPHA  /STATISTICS=DESCRIPTIVE SCALE CORR
 /SUMMARY=TOTAL MEANS VARIANCE CORR .
```

11. Notes additionnelles

a) Relation entre l'analyse factorielle, l'analyse en composantes principales et l'analyse de

fidélité.

Le but premier de ces analyses est d'en arriver à regrouper ensemble les items qui mesurent le même concept de façon à ce qu'une addition des réponses à un ensemble d'items constitue une nouvelle mesure, composite, d'un concept. Par exemple, si on additionne les réponses de chaque répondant à chacun des items ou questions mesurant la satisfaction envers un aspect extrinsèque de son travail, on obtiendra pour chaque répondant une mesure de la satisfaction extrinsèque. Cette mesure sera constituée de la somme – ou de la moyenne – de ses réponses aux diverses questions mesurant la satisfaction extrinsèque.

L'analyse en composantes principales (ACP) décompose la matrice de corrélation en tenant compte de l'ensemble de la variance des items. Elle en extrait un certain nombre de facteurs indépendants. *Le but de l'analyse en composantes principales est d'expliquer le plus de variance possible avec un nombre de composantes le plus restreint possible*. Après extraction, une part seulement de la variance totale est expliquée. Cette analyse donne un portrait global synthétique de la relation entre les variables.

L'analyse factorielle (AF) tient compte uniquement de la variance commune à l'ensemble des items. Elle extrait des facteurs qui peuvent être indépendants ou corrélés entre eux. *Son but est de reproduire le plus fidèlement possible la matrice de corrélation*. Comme l'ACP, elle permet de repérer des sous-ensembles de variables plus fortement corrélés entre eux.

Comme l'AF ne retient qu'une partie de la variance totale dans la solution finale, les résultats de l'analyse de fidélité peuvent contredire en partie ceux de l'analyse factorielle. On peut expliquer cette situation par le fait qu'il peut arriver que la variance commune de deux items sont bien corrélées mais que leurs variances spécifiques sont peu ou pas du tout corrélées ou même en corrélation négative. Comme l'analyse de fidélité considère l'ensemble de la variance, cette situation peut faire qu'un item bien identifié à un facteur en AF se révèle un mauvais contributeur à l'échelle.

b) Attention aux valeurs manquantes:

Comme il est possible que certains répondants n'aient pas répondu à tous les items d'un ensemble donné, le nombre de cas valides peut varier d'une analyse à l'autre selon que l'on retire ou que l'on ajoute un item. **Ceci peut modifier les résultats**. Il existe dans certaines procédures des moyens d'estimer les valeurs manquantes, entre autres en remplaçant la valeur manquante par la moyenne du groupe. **Comme règle générale, les valeurs manquantes ne sont pas estimées: les cas qui n'ont pas répondu à toutes les questions n'apparaissent pas dans l'analyse**. Nous n'avons pas le temps d'aborder toute cette problématique ici. Les logiciels proposent de plus en plus des méthodes d'imputation des valeurs manquantes, chacune ayant son biais propre. Il demeure qu'il faut examiner les résultats et la modification du nombre de cas valides selon les analyses et que le remplacement par la moyenne est habituellement tout à fait approprié lorsque la proportion de valeurs manquantes est faible (moins de 5%). Par ailleurs, au moment de créer des échelles additives, il est possible de calculer les moyennes pour tous les cas qui ont répondu

à un minimum d'items (de questions) mais pas à toutes. Ainsi, la fonction MEAN dans SPSS permet de calculer la moyenne et d'indiquer combien d'éléments au minimum doivent être répondus pour que le calcul soit valide.

Exemple: Pour une échelle ayant cinq variables, la commande suivante donnera la moyenne des réponses à cinq items ou questions si 5 items/questions sont répondus et la moyenne de quatre items si seulement quatre sont répondus. La valeur sera manquante pour les répondants ayant répondu à moins de quatre questions sur cinq.

Compute MESURE= MEAN (VI, V2, V3, V4, V5, 4).

Bibliographie

Tabachnik,Barbara G.; Fidell,Linda S. (2012): Using multivariate statistics. Harper and Row, New York. 983 pages.

Field, A. (2005). Discovering Statistics Using SPSS. London: Sage.779p.

Analyse factorielle et fidélité:

Carmines,Edward G.; Zeller,Richard A. (1979): Reliabiblity and validity assessment. (Quantitative applications in the social sciences, 17.) Sage, Beversly Hills. 71 pages.

Durand, C. et A. Blais, (2009) « La mesure », dans B. Gauthier, *Recherche sociale ; de la problématique à la collecte des données*, cinquième édition révisée, Québec, Presses de l'Université du Québec, p. 227-250.

Kim,Jae-On; Mueller,Charles (1978): Introduction to factor analysis. What it is and how to do it. (Quantitative applications in the social sciences.) Sage, Beverly Hills.

Kim,Jae-On; Mueller,Charles (1978): Factor analysis, statistical methods and practical issues. (Quantitative applications in the social sciences.) Sage, Beverly Hills.

Factor analysis and Principal components analysis, what's the difference?

by Lauri Tarkkonen

Factor analysis is based on the model

x = Bf+e, and cov(x)=Acov(f)A'+cov(e), one can make some further assumptions say cov(f)=I and cov(e)=D (diagonal), so the covariance is simplified to cov(x) = BB'+D = S.

Principal components are by definition:

u = A'x, such as V(u(i)) = a(i)'x is maximum a'a = 1 and the u(i) are orthogonal. Then cov(u) = A'cov(x)A, A'A and AA' are I, thus cov(x) = S = AA' . The factor analysis is a statistical model and the principal componentsare just a linear transformation of the original variables. Anyway, in factor analysis S-D = BB' and in principal components S = AA'.The difference in B and A is that you remove the D, the variances of the errors in FA but not in PCA. If there is no errors (it is easy to show that the same applies if all error variances are equal) the two methods will give you the same results.

If there is significant differences in the communalities of the variables, the two methods differ. So what is the proper one. If you do not assume measurement errors then use PCA; if you think there are measurement errors use FA.

I would like to put it this way:

> You all know the thing called the blueberry pie. First you take some dow, make a bottom of it, go to the forrest and pick some blueberries. If you believe in PCA, you put everything in your basket on top of the how. If you believe that there is something in your basket like leaves, needles, frogs that do not belong to the blueberry pie, you blow strongly while you pour your berries to the pie. Most things that do not belong to the pie fly away, you might loose some berries as well, but the taste is perhaps more of blueberries. If you did good job in picking your berries, there is no difference.

So why is this so difficult? Why do we always have this discussion?

First: Tradition.

In the beginning, there was T.W Anderson and his Introduction to Multivariate Analysis (-58). It started with PCA and told us that in the appendix (was it F) there is a funny thing called the Factor analysis. All the statisticians thought that PCA is the proper method and FA is some magic developed by some psychologists and you should not take it seriously.

Remember FA had this rotational indeterminacy.

Then: Lawley come up with the Maximum Likelihood solution. Now the statisticians had to accept FA as a bona fide statistical model. If there was Maximum Likelihood estimates for something, it must be the real thing. They realized that there was no theoretical base but the simplicity and the lack of indeterminacy speaking for the PCA.

More confusion:

Because both solutions were derived by the same solution of the eigenvalue problem, the spectral decomposition of a symmetric matrix, both analysis was performed with the same computer program. You just told it: do you have the communality estimation or not. Some programmers, like the maker of SYSTAT did not even understand the difference. Because the default value varied, so the naive users kept on getting whatever the programmers had to think as the main method.

Still more confusion:

The calculation of the factor solution has been two stage. Earlier the first stage was calculated by the 'principal axes' method. Some people do not see the difference with 'principal axes' and 'principal components'. They might have factors but they claim they have principal components.

Rotation

If you bother to stick to the definition of 'principal components' you will not rotate them, because if you rotate, the maximum variance part is not true anymore.

More tradition

In some areas the use of multivariate methods was started during the time when the statistician felt that PCA is the method and FA is some trick, there they swear still by PCA, because all the articles have PCA.

GREED

Some like PCA better because it gives sometimes higher loadings, but for some reason they still want to remove leaves and frogs from their blueberry pie.

Hope this helps.

- Lauri Tarkkonen

Lauri Tarkkonen / email: lauri.tarkkonen@helsinki.fi Tel:+358 0 666108
Korkeavuorenkatu 2 b B 11, 00140, Helsinki, Finland FAX +358 0 1913379