Analyse de données quantitatives (LPOLS1221)
Séance 4 : Analyse en composantes principales

1. Le coefficient de corrélation corrélation

Le coefficient de corrélation est une statistique qui permet de mesurer le degré de dépendance entre deux variables quantitatives. C'est-à-dire, que si la variable X augmente d'une unité, la variable Y augmentera ou diminuera de x unité. Le coefficient de corrélation se calcule sur la covariance entre les deux variables visées, divisé par le produit des deux écarts-types.

La covariance est une statistique permettant de connaître les écarts conjoints d'une observation vis-à-vis de deux variables ou plus. Son calcul est donc sensiblement différent à celui de la variance :

$$COV(x,y) = \frac{1}{n-1}\sum(x_i - \bar{X})(y_i - \bar{Y})$$

Et la corrélation se calculant de la manière suivante :

$$r = \frac{COV(x,y)}{S_x S_y}$$

Ce coefficient est donc compris entre -1 (association négative) et 1 (association positive). Un score proche de 0 renverra à une absence de relation. En sciences sociales, l'on a tendance à considérer les coefficients de corrélation selon la règle suivante :

- Autour de (-)0.1 : corrélation faible
- Autour de (-)0.3 : corrélation moyenne
- Autour de (-)0.5, au-dessus de 0.5 ou en-dessous de -0.5 : corrélation forte

Il n'en est bien sûr pas forcément de même dans d'autres domaines scientifiques.

2. L'analyse en composantes principales

L'analyse en composantes principales (ACP) est une analyse de réduction de l'information. Le principe est de trouver x dimensions égales au nombre de variables permettant de mieux expliqué la distribution des observations. Ces dimensions – appelées facteurs ou composantes - sont tracées par ordre de saturation d'information, ainsi le premier facteur expliquera toujours plus la dispersion des observations que le second et ainsi de suite. Cette portion d'explication de la variance des observations est également appelée en mathématiques des **valeurs propres**. La valeur propre est un chiffre compris entre 0 et le nombre de variables insérées dans l'analyse. La somme de toute les valeurs propres est égale au nombre de variables insérées dans l'analyse. Afin de réduire l'information, les chercheurs vont conserver un nombre limité de facteurs et ainsi diminué le nombre de variables utilisées. Il existe plusieurs méthodes pour sélectionner le nombre de facteurs à retenir dans l'analyse. Une règle d'or de retenir les facteurs dont la valeur propre est supérieure à 1. Sans quoi, nous ne faisons pas de la réduction d'information car nous préserverions des facteurs avec un potentiel d'explication inférieur à une 1 variable. Une autre règle est « le coude », lorsque l'on voit que les diminutions des valeurs propres ne sont plus fortement marquées, signifiant ainsi que la concentration de l'information devient moindre sur les axes suivants. Les facteurs sont des variables quantitatives centrées-réduites (moyenne de 0, écart-type de 1).

Il ne s'agit pas d'un test mais bel et bien d'un instrument multivarié à dessein descriptif, voire classificatoire. Dans un autre registre (plutôt celui de la psychologie), l'ACP peut également être utilisée pour valider des questionnaires ou des tests spécifiques.

Conditions d'application :

- Les **variables** doivent être **quantitatives.** Dans certains cas, il est possible d'inclure des variables **dichotomiques ou ordinales** mais, en aucun cas, nominales.
- L'ACP est adaptée lorsque les nombres d'observations et de variables sont élevés.
- Afin de réaliser une ACP, un **minimum de corrélation entre les variables** est nécessaire.

Les tests de **Kaiser-Meyer-Olkin** (KMO) et de **sphéricité de Bartlett** permettent de vérifier les corrélations minimales à l'acceptation des résultats de l'ACP.
Le *test KMO* donne un aperçu global des corrélations entre variables et se situe entre 0 et 1. Il est admis de considérer les seuils suivants pour l'interprétation du test :

- 0,80 et plus Excellent
- 0,70 et plus Bien
- 0,60 et plus Médiocre
- 0,50 et plus Misérable
- Moins de 0,50 Inacceptable

Le *test de sphéricité de Bartlett* quant à lui, renvoie une p-valeur. Si le test est significatif, on considère alors que les variables sont dépendantes les unes des autres et qu'une analyse en composante principale est possible (cela n'a pas de sens de réaliser une ACP si les variables sont indépendantes).

Autre méthode possible : analyser directement une matrice de corrélation.

3. Manipulations SPSS

Analyze → Dimension reduction → Factor

Dans l'encadré « Variable(s) », insérer la liste des variables numériques, dichotomiques ou ordinales sur lesquelles vous voulez effectuer l'ACP.

Dans les onglets suivants, sélectionnez :

Descriptives : Univariate descriptives, Initial solution, KMO and Bartlett's test of sphericity, signification levels (cette dernière est plutôt optionnelle : elle permet d'afficher la table des corrélations).

Extraction : méthod : principal components, unrotated factor solution & scree plot. On peut régler également manuellement les composantes principales à extraire selon la valeur propre que l'on désire.

Rotation : Varimax, rotated solution (/!\ très important pour l'interprétation)

Facteurs (scores): save as variable, regression

4. Remarques

Vérifiez toujours comment sont codées vos variables !! Pour simplifier l'interprétation des facteurs, vérifiez que les variables soient codées dans le même sens. Si ce n'est pas le cas, procéder à une modification des variables : « Compute variable » → Newvar = – (Oldvar).
Ne pas hésiter à demander des aides graphiques pour visualiser les composantes. Comme il s'agit de variables quantitatives, vous réaliserez des nuages de points (scatter dot) :

- Axe des Y : composante 1
- Axe des X : composante 2
- Option : Définir variable par : les modalités d'une variable particulière

Exercices

Chargez la base de données « European Values Study ».

1. Réalisez une analyse factorielle avec les variables ci-dessous et interpréter les résultats. Faites-le pour la population allemande.

Code question	Code variables	Intitulé variable
Q68a	V233	Do you justify : Claiming state benefits which you are not entitled to
Q68b	V234	Do you justify: Cheating on tax if you have the chance
Q68c	V235	Do you justify: Taking and driving away a car belonging to someone else (joyriding)
Q68d	V236	Do you justify: taking soft drugs
Q68e	V237	Do you justify: lying on own interest
Q68f	V238	Do you justify: Adultery
Q68g	V239	Do you justify: Accepting a bribe
Q68h	V240	Do you justify: Homosexuality
Q68i	V241	Do you justify: Abortion
Q68j	V242	Do you justify: Divorce
Q68k	V243	Do you justify: Euthanasia
Q68l	V244	Do you justify: Suicide
Q68m	V245	Do you justify: Paying cash to avoid taxes (Q68M)
Q68o	V247	Do you justify: Avoiding fare public transport (Q68O)
Q68q	V249	Do you justify: Experiments human embryos (Q68Q)
Q68r	V250	Do you justify: Manipulation food

2. Réalisez une analyse factorielle avec les variables ci-dessous et interpréter les résultats. Faites-le pour la population belge. En outre, produisez une statistique descriptive bivariée intéressante du facteur 1 avec la variable V114.

Code question	Code variables	Intitulé variable
Q42a	V136	Important in marriage: faithfulness
Q42b	V137	Important in marriage: adequate income
Q42c	V138	Important in marriage: same social background
Q42d	V139	Important in marriage: shared religious beliefs
Q42e	V140	Important in marriage: good housing
Q42f	V141	Important in marriage: agreement on politics
Q42i	V144	Important in marriage: share household chores
Q42j	V145	Important in marriage: children
Q42k	V146	Important in marriage: discuss problems
Q42l	V147	Important in marriage: time for friends and personal hobbies

3. Réalisez une analyse factorielle avec les variables ci-dessous et interpréter les résultats. Faites-le pour la population belge.

Code question	Code variables	Intitulé variable
Q63a	V205	how much confidence in: church
Q63b	V206	how much confidence in: armed forces
Q63c	V207	how much confidence in: education system
Q63d	V208	how much confidence in: the press
Q63e	V209	how much confidence in: trade unions
Q63f	V210	how much confidence in: the police
Q63g	V211	how much confidence in: parliament
Q63h	V212	how much confidence in: civil service
Q63i	V213	how much confidence in: social security system
Q63j	V214	how much confidence in: european union
Q63k	V215	how much confidence in: NATO
Q63l	V216	how much confidence in: united nations organisation
Q63m	V217	how much confidence in: health care system

Q63n	V218	how much confidence in: justice system
Q63o	V219	how much confidence in: major companies
Q63p	V220	how much confidence in: environmental organizations
Q63q	V221	how much confidence in: political parties
Q63r	V222	how much confidence in: government