PASW/SPSS : Analyse en composantes principales (ACP)

Vincent Jalby *

3 octobre 2009

`Analyse > Réduction des dimensions > Analyse factorielle`

1 Mise en œuvre

1.1 Descriptives

- **Statistiques - Caractéristiques univariées :** Donne l'espérance et l'écart-type de chaque variable.
- **Statistiques - Structure initiale :** Affiche la totalité de la solution (toutes les valeurs propres).
- **Matrice des corrélations - Coefficients :** Affiche la matrice des corrélations
- **Matrice des corrélations - Indice KMO et test de Bartlett :** Calcule le KMO et effectue le test de Bartlett. Ils permettent de déterminer, *a priori*, l'adéquation de l'ACP.
- **Matrice des corrélations - Anti-image :** Permet de déterminer les variables à supprimer dans le cas d'un KMO trop faible.

1.2 Extraction

- **Méthode :** Composantes principales correspond à l'ACP classique.
- **Analyser - Matrice de corrélation/covariance :** fait une ACP normée ou non-normée.
- **Afficher - Structure factorielle sans rotation :** Résultat avant rotation [Laisser cocher – supprimer éventuellement si rotation]. Affiche les coordonnées des composantes, …
- **Afficher - Diagramme des valeurs propres :** Scree plot. Permet de déterminer le nombre de composantes à retenir en repérant le *coude*.
- **Extraire :** Permet de préciser le nombre de composantes (ou facteurs) que l'on souhaite retenir :
  - **Basé sur la valeur propre :** La règle de Kaiser ne retient que les composantes dont la valeur propre est supérieure à 1 (réglage par défaut).
  - **Extraire - Nombre de facteurs :** Nombre de composante à retenir.

1.3 Rotation

- **Méthode - Aucune/Varimax/Quartimax/Equamax :** Effectue une rotation dans le plan factoriel. Ne change pas l'inertie expliquée par le plan.
  - **Aucune :** Pas de rotation. Les composantes correspondent aux valeurs propres par ordre de grandeur décroissante.
  - **Varimax :** simplifie l'interprétation des composantes
  - **Quartimax :** simplifie l'interprétation des variables
  - **Equamax :** combinaison de Varimax et Quartimax
- **Afficher - Structure après rotation :** Affiche les coordonnées des composantes après rotation, …
- **Afficher - Carte(s) factorielle(s) :** Graphique des variables dans le plan factoriel.

1.4 Facteurs

- **Enregistrer dans des variables (Méthode Régression) :** Permet d'enregistrer (les coordonnées des individus dans) les nouvelles variables. Permet de faire une représentation du nuage des individus.
- **Afficher la matrice des coefficients factoriels :** Coordonnées des composantes dans les variables initiales.

2 Résultats

2.1 Statistiques descriptives

Affiche la moyenne, l'écart-type et le nombre d'observations pour chaque variable. Permet donc de

* Faculté de Droit et de Sciences Économiques, Université de Limoges. E-mail: `vincent.jalby@unilim.fr`

– juger de l'hétérogénéité des variables ;
– *repérer* les variables ayant des valeurs manquantes.

2.2 Matrices de corrélation

Permet de déceler rapidement les variables fortement corrélées et/ou de juger de l'existence de corrélations suffisantes entre les variables. À confirmer par le *test de Bartlett*.

2.3 Test de sphéricité de Bartlett

Ce test consiste à comparer la matrice des corrélations $X'X$ avec l'identité (pas de corrélation entre les variables) en utilisant un test du χ^2. Une valeur élevée avec une signification proche de 0 permet de rejeter la non-corrélation globale des variables, c'est-à-dire, assure que les variables sont suffisamment corrélées entre-elles pour permettre une réduction significative de la dimension. Condition indispensable pour faire une ACP.

2.4 Test Kaiser-Mayer-Olkin

Le KMO, rapport de la somme des corrélations au carré par la somme des corrélations partielles au carré, est un réel compris entre 0 et 1. Un KMO assez élevé (> 0.6) assure que les corrélations partielles ne sont pas trop importantes par rapport aux corrélations *simples*. Indispensable pour obtenir une ACP intéressante. Dans la négative, il peut être nécessaire de supprimer certaines variables.

2.5 Graphique des valeurs propres

Repérer dans le Scree plot, le « coude » des valeurs propres. Il faudrait retenir toutes les valeurs propres (et donc les composantes associées) jusqu'au coude.

2.6 Qualité de représentation

Repérer les variables ayant un taux d'extraction (de variance) faible, en dessous de 60 %. L'interprétation de ces variables devra être faite avec prudence. Cette étape peut être une confirmation des observations faites sur le graphe.

2.7 Variance totale expliquée

Déterminer le nombre de composantes à retenir pour avoir plus de 70 % de variance (cumulée) expliquée. Si le nombre de composantes est supérieur à 2, il faudra étudier plusieurs schémas. L'importance de chaque composante est donnée par le % de variance expliquée (par chaque composante).

2.8 Matrice des composantes (après rotation)

Coordonnées des variables dans les composantes.

2.9 Matrice de transformation

Rotation des composantes par rapport aux composantes principales théoriques.

2.10 Matrice des coefficients des coordonnées des composantes

Coordonnées des composantes dans les variables initiales.

2.11 Matrice des covariances des composantes

Identité car orthogonales (non corrélées).

3 Analyse de l'ACP

3.1 Intérêt de l'ACP : KMO and Bartlett's Test / Correlation Matrix

Vérifier que le Chi-2 du Bartlett's Test est suffisamment grand avec une signification quasi nulle : les variables sont suffisamment corrélées. La matrice des corrélation peut confirmer cela. Vérifier que le KMO est supérieur à 0,6 ou 0,5 : pas de corrélations partielles trop importantes. Sinon, supprimer une ou plusieurs variables de l'analyse.

3.2 Qualité de l'ACP : Variance totale expliquée / Graphique des valeurs propres

Déterminer le nombre de composantes à retenir pour avoir plus de 70 % de variance (cumulée) expliquée. Si le nombre de composantes est supérieur à 2, il faudra étudier plusieurs schémas. L'importance de chaque composante est donnée par le % de variance expliquée (par chaque composante). Repérer dans le Scree plot, le « coude » des valeurs propres. Il faudrait retenir toutes les valeurs propres (et donc les composantes associés) jusqu'au coude. Cela doit correspondre au nombre de composantes déterminé précédemment.

3.3 Qualité de représentation des variables : Qualité de représentation

Repérer les variables ayant un taux d'extraction (de variance) faible, en dessous de 60 %. L'interprétation de ces variables devra être faite avec prudence. Cette étape peut être une confirmation des observations faites sur le graphe.

3.4 Interprétation des composantes / Contribution des variables : Matrice des composantes

Repérer les variables ayant une forte contribution (positive ou négative) sur chaque composante. Ces variables donneront un sens aux composantes. Deux (groupes de) variables avec des contributions de signes opposés représenteront des oppositions. Cette étape peut être une confirmation des observations faites sur le graphe.

3.5 Interprétation graphique : Diagramme des composantes

L'étude graphique ne doit porter que sur les variables se trouvant proches du cercle (bord du disque) des corrélations, c'est-à-dire celles qui sont suffisamment représentées. Repérer les groupes de variables et interpréter leurs regroupements. Des variables proches représentent des variables fortement corrélées. Des variables « à angle droit » représentent des variables non corrélées. Les variables proches des axes permettent de donner un sens aux composantes, en mettant éventuellement en valeur des oppositions.

4 Nuage des individus

4.1 Coordonnées des individus

Pour obtenir le nuage des individus (dans le plan factoriel), il faut faire une ACP en ayant coché l'option `Facteurs > Enregistrer dans des variables`. Deux (ou plus) nouvelles variables sont générées. Elles portent le nom `facx_y` où `x` représente le numéro du facteur, et `y` le numéro de l'analyse.

4.2 Diagramme des individus

Faire alors un diagramme de dispersion simple (`Graphe > Boites... > Dispersion/Points > Dispersion simple`). Mettre le premier facteur sur l'axe X et le second sur l'axe Y. Étiqueter les observations par la variable contenant le nom des individus, et ne pas oublier de cocher dans `Options...` l'option `Afficher le diagramme avec les étiquettes d'observations`.

4.3 Interprétation du nuage des individus

L'origine des axes $(0, 0)$ correspond à la moyenne sur l'échantillon. La signification des axes est celle obtenue dans l'analyse duale (des variables).
Il convient de mettre en valeur :
– les groupes d'individus (ayant donc un comportement identique) ;
– les individus isolés ;
– la position relative des (groupe d') individus par rapport aux axes.
Attention, ce graphique ne permet pas de connaître la qualité de représentation des individus.

4.4 Contribution

La contribution d'un individu X_i à la détermination de l'axe U_λ est donnée par

$$\mathrm{CTR}_\lambda(i) = \frac{m_i F_\lambda^2(i)}{\lambda}$$

Les points les plus éloignés de l'origine ont les plus fortes contributions.

4.5 Qualité de représentation des individus

Il n'est pas possible de l'obtenir automatiquement. Les formules théoriques sont :

$$\mathrm{QLT}(i) = \sum_{\lambda} \mathrm{CO2}_{\lambda}(i) \qquad \mathrm{CO2}_{\lambda}(i) = \frac{F_{\lambda}^2(i)}{\|X_i - G\|^2}$$

où λ représente les valeurs propres des composantes retenus, X_i l'individu i, G le barycentre des individus, $F_{\lambda}(i)$ la coordonnée de X_i sur l'axe associé à λ, $\mathrm{CO2}_{\lambda}(i)$ le taux de représentation de X_i par l'axe associé à λ, $\mathrm{QLT}(i)$ la qualité de représentation de X_i dans les axes associés aux λ.
Pour appliquer ces formules dans SPSS, il faut tenir compte que
– les calculs sont faits sur des données centrées-réduites ($X_i = X_i', G = 0$)
– les coordonnées données par SPSS (`facx_y`) sont données dans un système d'axes orthonormaux.
Pour appliquer les formules précédentes, il faut donc centrer et réduire les variables originales et multiplier les coordonnées sur les axes principaux par $\sqrt{\lambda}$.

4.5.1 Normalisation des variables

Utiliser `Analyse > Statistiques descriptives > Descriptives` en cochant `enregistrer des valeurs standardisées dans des variables` sur les variables originales.

4.5.2 Norme de chaque point

Définir une nouvelle variables `norm2` via `Transformer > Calculer` en utilisant la formule :

$$\texttt{norm2} = \texttt{z\_variable\_1} * *2 + \cdots + \texttt{z\_variable\_2} * *2$$

4.5.3 Calcul des CO2

Définir les nouvelles variables `CO2_1`, `CO2_2` pour chacun des axes via `Transformer > Calculer` en utilisant la formule :

$$\texttt{CO2\_-} = (\texttt{fac-\_1} * *2) * \lambda / \texttt{norm2}$$

4.5.4 Calcul de QLT

Définir une nouvelle variable `QLT`, via `Transformer > Calculer` en utilisant la formule :

$$\texttt{QLT} = \texttt{CO2\_1} + \texttt{CO2\_2} + \ldots$$

4.5.5 Cas de la rotation

En cas de rotation, il n'est pas possible d'utiliser les formules précédentes.

> En effet, la rotation est effectuée dans l'espace des variables ; lorsqu'on l'applique à l'espace des individus, il s'agit d'une rotation composée avec une homotéthie sur chacune des variables.
> Soit (f_1, f_2) les coordonnées d'un individus X dans les axes factoriels avant rotation, λ_1, λ_2 les valeurs propres associées à chaque axe. Alors les coordonnées de cet individus dans l'espace des individus sont $(x_1, x_2) = (\sqrt{\lambda_1} f_1, \sqrt{\lambda_2} f_2)$.
> Soit $R = \begin{pmatrix} a & b \\ b & -a \end{pmatrix}$ la matrice de rotation. Après rotation, les coordonnées de l'individu X sur les axes factoriels sont $(f_1', f_2') = (af_1 + bf_2, bf_1 - af_2)$, mais ses coordonnées dans l'espace des individus sont $(x_1', x_2') = (a\sqrt{\lambda_1} f_1 + b\sqrt{\lambda_2} f_2, b\sqrt{\lambda_1} f_1 - a\sqrt{\lambda_2} f_2)$. Il n'existe pas d'expression simple de (x_1', x_2') en fonction de (f_1', f_2').
> Les CO2 après rotation sont donc (pour la première composante)
>
> $$\mathrm{CO2\_R}_1 = \frac{|x_1'|^2}{\|X\|^2} = \frac{a^2 f_1^2 \lambda_1 + b^2 f_2^2 \lambda_2}{\|X\|^2}$$
>
> Il n'est donc pas possible d'exprimer simplement les CO2 après rotation en fonction de (f_1', f_2').

En dimension 2, on peut facilement déduire les CO2 après rotation de ceux avant rotation via les formules :

$$\texttt{CO2\_1\_R} = a^2 \texttt{CO2\_1} + b^2 \texttt{CO2\_2} \quad \text{et} \quad \texttt{CO2\_2\_R} = b^2 \texttt{CO2\_1} + a^2 \texttt{CO2\_2}$$

où a et b sont les coefficients (des colonnes) de la matrice de rotation. Bien sûr, les `QLT` ne changent pas. (Ces formules se généralisent simplement aux dimensions supérieures, en lisant en colonne les coefficients de la matrice de rotation.)

5 Amélioration de l'ACP

5.1 Rotation

Si l'interprétation des composantes n'est pas convaincante, utilisez une rotation pour obtenir une nouvelle analyse :
- **Varimax :** simplifie l'interprétation des composantes en minimisant le nombre de variables ayant de fortes contributions sur une même composante
- **Quartimax :** simplifie l'interprétation des variables en minimisant le nombre de composantes nécessaires à l'explication de chaque variable
- **Equamax :** compromis entre Varimax et Quartimax.

5.2 Suppression de variables

5.2.1 Test de Bartlett

Si le test de Bartlett échoue (variables insuffisamment corrélées), il y a peu d'espoir d'améliorer l'ACP.

5.2.2 Amélioration du KMO

Si l'indice KMO est trop faible (< 0.5), cela signifie qu'il y a trop de corrélations partielles. Il convient donc de supprimer la (ou les) variables ayant le plus d'*influence* sur les corrélations partielles. Pour cela, demander le calcul de la matrice des corrélations « anti-image ». La diagonale de cette matrice correspond au KMO pour chaque variable (quotient de la somme des corrélations au carré de cette variable avec les autres variables, par la même chose plus la somme des corrélations partielles au carré de cette variable.) Il convient donc de supprimer la variable ayant le KMO le plus faible.

5.2.3 Contributions excessives

Si une variable (ou un individu) a une contribution trop importante sur (la détermination d') une composante principale, il peut être intéressant de supprimer cette variable (ou cet individu) de l'étude pour tenter de mieux expliquer les autres variables.