ANALYSE EN FACTEURS COMMUNS ET SPÉCIFIQUES (I, K5)
(07 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

L'**analyse en facteurs communs et spécifiques** (afcs) est une méthode d'**analyse des données**, issue de la psychologie, dont l'objet consiste à déterminer, à partir de **corrélations** empiriques entre des variables données, des **facteurs « cachés »**, en nombre minimum, susceptibles d' « expliquer » au mieux ces variables initiales.

(i) Dans l'**analyse théorique**, on considère un **modèle statistique** $(\Omega, \mathcal{T}, \mathcal{P})$ et un **vecteur aléatoire observable** $\xi : \Omega \mapsto \mathbf{R}^K$. On suppose qu'il existe un vecteur $\varphi = (\varphi_1 ,..., \varphi_H)'$ (avec $H < K$) et un vecteur $\varepsilon = (\varepsilon_1 ,..., \varepsilon_K)'$ tous deux tq ξ décompose linéairement en fonction de (φ, ε) selon :

(1) $\quad \xi = A\,\varphi + D\,\varepsilon,$

où $A \in M_{KH}(\mathbf{R})$ et $D \in D_K(\mathbf{R})$ (**matrice diagonale**). Pour tout $k \in N_K^*$, la k-ième ligne de l'équation vectorielle (1) s'explicite selon :

(2) $\quad \xi_k = \sum_{h=1}^{H} a_{kh}\,\varphi_h + d_k\,\varepsilon_k ,$

et reçoit l'interprétation suivante :

(a) φ_h est, $\forall\, h \in N_H^*$, un **facteur commun** à tous les ξ_k . Il existe donc $H < K$ facteurs communs ;

(b) ε_k est, $\forall\, k \in N_K^*$, un **facteur spécifique** de ξ_k ;

(c) les paramètres a_{kh} sont appelés **coefficients factoriels communs**, ou **poids factoriels communs**.

En psychométrie, ils sont aussi appelés **coefficients de saturation des facteurs communs**, ou **coefficients de test des facteurs communs**. La matrice $A = (a_{kh})_{(k,h)}$ est appelée **matrice des coefficients des facteurs communs** (en abrégé, **matrice des communalités**, ou **matrice commune**).

Certains coefficients de A peuvent être posés comme nuls a priori : la matrice A, dans laquelle ces **restrictions a priori** sont supposées présentes, décrit donc un **schéma factoriel** dont le rôle est de préciser les facteurs réellement communs à tous les ξ_k et les facteurs communs seulement à certaines variables ξ_k (**facteurs de groupes**) ;

(d) les paramètres d_k sont appelés **coefficients factoriels spécifiques**, ou **poids factoriels spécifiques**. En psychologie, on les appelle aussi **coefficients de saturation des facteurs spécifiques**, ou **coefficients de test des facteurs spécifiques**. La **matrice diagonale** $D = (d_k)_k$ est appelée **matrice des coefficients des facteurs spécifiques** (en abrégé **matrice des spécificités**, ou **matrice spécifique**).

Si ξ est de carré intégrable (ie $\xi \in L_{RK}^2 (\Omega, \mathcal{T}, P)$, $\forall P \in \mathcal{P}$), sa **dispersion** s'écrit (les matrices A et D étant supposées non aléatoires) :

(2) $\quad V \xi = A (V \varphi) A' + A C (\varphi, \varepsilon) D' + D C (\varepsilon, \varphi) A' + D (V \varepsilon) D'$,

où $V \varphi = C (\varphi, \varphi) = E (\varphi - E \varphi)(\varphi - E \varphi)'$ est la **matrice de dispersion** de φ, $V \varepsilon$ celle de ε et $C (\varphi, \varepsilon) = E (\varphi - E \varphi)(\varepsilon - E \varepsilon)'$ la **matrice des covariances** croisées de φ et ε, avec $C (\varepsilon, \varphi) = C (\varphi, \varepsilon)'$ (cf **covariance croisée**).

Si les facteurs φ et ε sont corrélés entre eux, on les appelle des **facteurs obliques**, et l'on définit ainsi la notion d'**analyse factorielle oblique** (ie dans laquelle $C (\varphi, \varepsilon) \neq O_{KH}$.

Si les facteurs φ et ε ne sont pas corrélés entre eux, on définit l'**analyse factorielle simple**, ou **analyse factorielle orthogonale**. Dans ce cas, la k-ième ligne de l'équation vectorielle (2) s'écrit, $\forall k \in N_K^*$:

(2)' $\quad V \xi_k = \Sigma_{h=1}^{H} a_{kh}^2 V \varphi_h + d_k^2 V \varepsilon_k$.

Le premier terme est appelé **variance des facteurs communs** (ou **communalité**) et le second **variance des facteurs spécifiques** (ou **spécificité**). On appelle de façon analogue les matrices $A (V \varphi) A'$ (**dispersion commune**) et $D (V \varepsilon) D'$ (**dispersion spécifique**).

(ii) Le cadre précédent définit l'**analyse en facteurs communs et spécifiques**, qui fait partie des méthodes d'**analyse multidimensionnelle** (les anglo-saxons l'appellent simplement *« analyse factorielle »*).

Une variante du schéma précédent distingue entre facteurs proprement spécifiques et **perturbation aléatoire** (sur l'équation). La forme (1) est alors remplacée par la forme :

(1)" $\quad \xi = A \varphi + B \psi + C \varepsilon$,

dans laquelle φ est le vecteur des facteurs communs φ_h , ψ celui des facteurs spécifiques ψ_k , et ε celui des perturbations aléatoires (appelées **incertitudes** en psychologie) ε_k . La terminologie, la démarche et les formules précédentes s'adaptent au couple (φ, ψ), qui remplace le couple (φ, ε).

(iii) En l'absence de **contraintes a priori** sur les paramètres, le modèle précédent n'est pas identifiable (cf **identification**, **modèle identifiable**). De plus, on doit associer au modèle (1) des hypothèses stochastiques, dont les plus courantes sont les suivantes :

(a) les variables sont centrées (ie $E \varphi = 0$ et $E \varepsilon = 0$, d'où $E \xi = 0$) ;

(b) les variables sont sans corrélation (ie $C (\varphi, \varepsilon) = 0_{KH}$) ;

(c) la dispersion de φ est unitaire (ie $V\,\varphi = I_H$).

Des hypothèses plus générales entraînent, en effet, une analyse plus complexe (cf analyse oblique précédente).

(iv) En pratique, l'analyse empirique dispose d'**observations** $X_n' = (x_{n1}, ..., x_{nK})$ ($\forall$ n $\in N_N^*$) du vecteur ξ. On note $X = (x_{nk})_{(n,k)} \in M_{NK}$ (**R**) la **matrice des observations** ainsi définie. On note, de façon analogue, $F_n' = (f_{n1}, ..., f_{nH})$ les valeurs (**inobservables**) correspondant à φ et $U_n' = (u_{n1}, ..., u_{nK})$ les valeurs (inobservables) correspondant à ε. On pose $F = (f_{nh})_{(n,h)} \in M_{NH}$ (**R**) ainsi que $U = (u_{nk})_{(n,k)} \in M_{NK}$ (**R**).

Par suite, le modèle (1) est « observé » selon la forme :

(3) $X_n' = A\,F_n' + D\,U_n'$,

ou encore, sous forme matricielle :

(4) $X' = A\,F' + D\,U'$,
$\quad\;\; (K,N)\;\; (K,H)\;\; (H,N)\;\; (K,K)\;\; (K,N)$

expression parfois écrite sous forme transposée $X = F\,A' + U\,D$ (car $D' = D$).

Par suite, sous les hypothèses classiques suivantes :

(a) X est centrée pr à sa moyenne. Autrement dit, si $T \in M_{NK}$ (**R**) est le **tableau statistique** initial, on a $X = P\,T$, où P désigne la **matrice de centrage par rapport à la moyenne** ;

(b) F et U sont centrées pr à leurs moyennes respectives ;

(c) F possède une dispersion « empirique » unitaire, au sens où l'on a $V\,F = N^{-1}\,F'\,F = I_H$, et U une dispersion « empirique » diagonale, au sens où $V\,U = N^{-1}\,U'\,U \in D_K$ (**R**) (matrice diagonale) ;

(d) F et U sont sans corrélation « empirique », ie $C_{FU} = N^{-1}\,F'\,U = 0_{HK}$;

(e) $D = I_K$,

on obtient une formule, analogue à (2), relative à la dispersion empirique de X :

(5) $X'\,X = A\,F'\,F\,A' + D\,U'\,F\,A' + A\,F'\,U\,D + D\,U'\,U\,D$.

Cette dernière se simplifie selon :

(6) $V\,X = N^{-1}\,X'\,X = A\,A' + V_U$,

où A est la matrice (inobservable) des coefficients des facteurs communs et V_U la matrice des covariances empiriques (inobservable) des facteurs spécifiques ε_k (matrice dont la diagonale comporte les variances spécifiques).

(v) L'analyse en facteurs communs et spécifiques est liée à l'**analyse en composantes principales** (acp) (centrée). Cette dernière revient ici, à diagonaliser la matrice des covariances empiriques $V_X = N^{-1} X' X$ à l'aide d'une matrice orthogonale Q, ie à utiliser la **décomposition spectrale** suivante :

(7) $\quad V_X = Q \Lambda Q' = Z Z', \qquad$ avec $Z = Q \Lambda^{1/2}$,

dans laquelle on suppose que les colonnes de Q sont les vecteurs propres, rangés selon l'ordre décroissant des K valeurs propres $\lambda_1 \geq ... \geq \lambda_{[rg\ Q]} \geq 0 \geq ... \geq 0$ formant la diagonale de la matrice diagonale Λ (où rg Q désigne le rang de Q).

Dans l'analyse en facteurs communs et spécifiques, le modèle résumé en (6) revient simplement à supposer que c'est la matrice $V_X - V_U$ qui se décompose sous la forme A A'. Néanmoins, si les variances figurant sur la diagonale de V_U sont « petites », les résultats obtenus par les deux méthodes sont voisins.

(vi) Pour estimer alors le modèle de l'analyse factorielle (1), on cherche généralement à estimer la forme (6), ie à estimer A et V_U (soit (H+1) . K paramètres). Des procédures existent, eg les suivantes :

(a) à distance finie (ie si $(H + 1) . K < N << +\infty$), si l'on peut estimer la matrice V_U par une matrice $V_U\tilde{}$, on applique l'**acp** (centrée) à la matrice $V_X - V_U\tilde{}$. L'unicité de la solution du problème initial n'est cependant pas assurée ;

(b) toujours à distance finie, et si les variances spécifiques sont égales entre elles (ie si $V U = \sigma_U^2 I_K$), une procédure itérative est la suivante :

$(b)_0$ on pose $V_U^{(0)} = 0$ et l'on applique l'acp à V_X , d'où $V_X = Z Z'$. On retient les H premiers facteurs de Z, constituant une matrice $Z^{(1)}$;

$(b)_1$ on calcule $V_U^{(1)} = V_X - Z^{(1)} Z^{(1)}{}'$ et l'on applique l'acp à la matrice $V_X - V_U^{(1)}$, d'où une décomposition $V_X - V_U^{(1)} = Z_1 Z_1'$. On retient les H premiers facteurs de Z_1 , qui constituent une matrice $Z^{(2)}$;

$(b)_2$ on calcule $V_U^{(2)} = V_X - Z^{(2)} Z^{(2)}{}'$, etc.

(c) sous certaines conditions (eg la **normalité** des variables, prise en compte des restrictions a priori sur A et **identifiabilité** du modèle (1)), on peut appliquer la **méthode du maximum de vraisemblance**.

(vii) Enfin, si l'on suppose A estimée par une matrice $A\tilde{}$ et D par une matrice $D\tilde{}$, on peut tester la significativité des H facteurs communs (ie l'hypothèse $H_0 : A = 0_{KH}$) à l'aide de la statistique :

(8) $T_N = (N - 1) . \text{Log} \{|\tilde{A} \tilde{A}'|^{-1} . |\tilde{A} \tilde{A}' + \tilde{D}^2|\}$,

qui possède (sous certaines hypothèses) des propriétés de **convergence en loi** tq la suivante :

(9) $\mathcal{L}(T_N) \rightarrow_{N \rightarrow +\infty} \mathcal{X}_d^2$ (**loi du chi-deux** à d **degrés de liberté**),

où $d = [(1/2) \{(N-H)^2 + (N-H)\}]$ et où [.] désigne la fonction **partie entière**.