

ANALYSE EN COMPOSANTES PRINCIPALES (I, K5)

(08 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

L'analyse en composantes principales est une méthode classique d'analyse des données qui se ramène à l'analyse générale des données par transformation du tableau statistique initial (cf transformation des données). Cette analyse vise à « remplacer » un ensemble de variables par un autre ensemble de variables, non corrélées entre elles (les « composantes principales ») et en nombre moindre que celui des variables initiales.

On présente trois variantes de ce type d'analyse, selon qu'elle met en oeuvre un « tableau brut » (ie non transformé), un « tableau centré » ou un « tableau normé ».

(i) On considère un modèle statistique de base $(\Omega, \mathcal{F}, \mathcal{P})$ et un vecteur aléatoire réel $\tau : \Omega \mapsto \mathbf{R}^K$, supposé de carré intégrable (ie, $\tau \in L^2_{\mathbf{R}^K}(\Omega, \mathcal{F}, P)$, $\forall P \in \mathcal{P}$). On dispose de N observations (non nécessairement indépendantes) $T_n' = (t_{n1}, \dots, t_{nK})$ (avec $n \in N_N^*$) du vecteur τ et l'on note $T = (t_{nk})_{(n,k)} \in M_{NK}(\mathbf{R})$ la matrice des observations, ou tableau statistique « brut » initial, qui en résulte :

$$(1) \quad T = (t_{nk})_{(n,k)} = [T_1 \dots T_N]' = [t^1, \dots, t^K].$$

où les T_n sont des lignes à valeurs dans $\mathbf{R}^K$, et les t_k des colonnes à valeurs dans $\mathbf{R}^N$.

On définit alors les caractéristiques théoriques suivantes :

(a) l'espérance $\mu = E \tau$, la matrice de dispersion $\Sigma = V \tau$ et la matrice de corrélation $\Gamma = C \tau$ de τ , avec $\Sigma = (\sigma_{kl})_{(k,l)}$, matrice dont les éléments sont les covariances, et $\Gamma = (\gamma_{kl})_{(k,l)}$, avec $\gamma_{kl} = C(\tau_k, \tau_l) (\sigma_{kk} \sigma_{ll})^{-1/2} = \gamma_{kl} = \sigma_{kl} (\sigma_{kk} \sigma_{ll})^{-1/2}$ (coefficients de corrélation linéaire), où σ_{kl} est la covariance, terme général de Σ ;

(b) la matrice de dispersion $\sigma = \text{diag} \{\sigma_{11}, \dots, \sigma_{KK}\}$ (matrice diagonale), dont les termes diagonaux sont ceux de Σ . Par suite, $\Gamma = \sigma^{-1/2} \Sigma \sigma^{-1/2}$;

(c) le vecteur centré $\xi = \tau - \mu$, qui vérifie donc (cf variable centrée) : $E \xi = 0$ et $V \xi = \Sigma$;

(d) le vecteur normé (ou vecteur centré-réduit, ou encore vecteur standardisé) $\varepsilon = \sigma^{-1/2} \xi = \sigma^{-1/2} (\tau - \mu)$, qui vérifie (cf variable réduite) : $E \varepsilon = 0$ et $V \varepsilon = \Gamma$.

On définit aussi les analogues empiriques suivants :

(a) la moyenne empirique vectorielle $\bar{t} = (\bar{t}_1, \dots, \bar{t}_K)'$, avec $\bar{t}_k = N^{-1} \sum_{n=1}^N t_{nk}$, (moyenne des t_k), $\forall k$;

(b) la **matrice de dispersion** (ou des variances-covariances) empirique (cf **matrice de covariance**) :

$$(2) \quad V_T = T' P T / e_N' e_N ,$$

(où P est la **matrice de centrage par rapport à la moyenne**), dont les éléments sont $v_{kl} = N^{-1} \sum_{n=1}^N (t_{nk} - \bar{t}_k)(t_{nl} - \bar{t}_l) = t_k' P t_l / e_N' e_N$ (covariances empiriques) ;

(c) la **matrice des corrélations** :

$$(3) \quad C_T = (c_{kl})_{(k,l)} , \quad \text{avec } c_{kl} = v_{kl} (v_{kk} v_{ll})^{-1/2} .$$

Ces trois notions empiriques (moyenne, dispersion et corrélation) sont donc ainsi associées à T. On note souvent $S = (s_{kl})_{(k,l)}$ ou S_N la dispersion V_T précédente ;

(d) la **matrice diagonale** de dispersion $s = \text{diag} \{v_{11}, \dots, v_{KK}\}$, dont les termes diagonaux sont ceux de V_T : ie $v_{kk} = \|P t_k\|^2 / e_N' e_N$, $\forall k$. Par suite, on a :

$$(3)' \quad C_T = s^{-1/2} V_T s^{-1/2}$$

(ou $s^{-1/2} S s^{-1/2}$, ou encore $s^{-1/2} S_N s^{-1/2}$ selon les notations retenues) ;

(e) la **matrice centrée** $X = P T$, tq $x_k = t_k - \bar{t}_k e_N$ ($\forall k$). Cette matrice est donc tq $e_N' X = 0$ et $N^{-1} X' X = T' P T = V_T$ (**dispersion** empirique) ;

(f) la matrice normée $U = X s^{-1/2} = P T s^{-1/2}$ Cette matrice est donc tq $e_N' U = 0$ et $N^{-1} U' U = s^{-1/2} T' P T s^{-1/2} = s^{-1/2} V_T s^{-1/2} = C_T$.

(ii) L'**analyse théorique** peut alors mettre en oeuvre :

(a) l'analyse en composantes principales **simple** (ou élémentaire) du vecteur τ : ie l'étude (ie la **décomposition spectrale**, aussi appelée décomposition en valeurs singulières) de sa matrice des moments (non centrés) ;

(b) l'analyse en composantes principales **centrée** de τ : ie l'étude (décomposition spectrale) de sa matrice de dispersion Σ ;

(c) l'analyse en composantes principales **normée** (ou **centrée-réduite**, ou encore **standardisée**) de τ : ie l'étude (décomposition spectrale) de sa matrice des corrélations Γ .

En pratique, l'**analyse empirique** utilise le tableau T et ses transformés. On appelle alors :

(a) **analyse en composantes principales simple** (ou élémentaire) de T l'application de l'analyse générale des données au tableau T lui-même ;

(b) **analyse en composantes principales centrée** de T l'application de l'analyse générale des données au tableau centré $X = P T$. La matrice à diagonaliser, $X' X$, est donc proportionnelle à la matrice des covariances empirique entre variables (ie $X' X = N V_T$) ;

(c) **analyse en composantes principales normée** (ou centrée-réduite, ou encore standardisée) de T l'application de l'analyse générale des données au tableau normé $U = P T s^{-1/2}$. La matrice à diagonaliser, $U' U$, est donc proportionnelle à la matrice des corrélations empirique entre variables (ie $U' U = N C_T$).

(iii) Dans le cas de l'**analyse centrée**, on note :

$$(4) \quad \Sigma = Q \Lambda Q' = \sum_{k=1}^K \lambda_k q_k q_k'$$

la **décomposition en valeurs singulières** de Σ , où $Q = [q_1, \dots, q_K]$ est la **matrice orthogonale** (de changement de **base**) qui diagonalise Σ et où $\Lambda = \text{Diag} \{\lambda_1, \dots, \lambda_K\}$ est la matrice diagonale correspondante, supposée tq $\lambda_1 \geq \dots \geq \lambda_K \geq 0$. Pour tout k , on appelle alors k -ième **composante principale** de τ (ou de ξ) la **vars** :

$$(5) \quad \gamma_k = q_k' (\tau - \mu) = q_k' \xi.$$

Le vecteur des composantes principales γ est défini selon :

$$(6) \quad \gamma = (\gamma_1, \dots, \gamma_K)' = Q' (\tau - \mu) = Q' \xi.$$

On note de même :

$$(4)' \quad V_T = Z L Z' = \sum_{k=1}^K l_k z_k z_k'$$

la décomposition en valeurs singulières de V_T , où $Z = [z_1, \dots, z_K]$ est la matrice orthogonale (de changement de base) qui diagonalise V_T et $L = \text{Diag} \{l_1, \dots, l_K\}$ la matrice diagonale supposée tq $l_1 \geq \dots \geq l_K \geq 0$ (certains termes diagonaux pouvant être nuls).

Pour tout $l = 1, \dots, L$, on appelle alors k -ième **composante principale** de t_l (ou de x_l) la **vars** :

$$(5)' \quad c_{kl} = z_l' (t_k - \bar{t}_k e_N) = z_l' x_k.$$

D'où le vecteur des composantes principales :

$$(6)' \quad c(k) = (c_{k1}, \dots, c_{kK})' = Z' (t_k - \bar{t}_k e_N) = Z' x_k,$$

vecteur qui définit la **matrice des composantes principales** :

$$(6)'' \quad C = [c(1), \dots, c(K)].$$

Si les vecteurs initiaux T_n' sont indépendants et équidistribués selon $T_n' \sim \mathcal{N}_K(\mu, \Sigma)$ (cf **suite iid**), on établit, sous certaines conditions, la **normalité asymptotique** suivante (**convergence en loi**) :

$$(7) \quad N^{1/2} (l_k - \lambda_k) \rightarrow_{N \rightarrow \infty} \mathcal{N}_1(0, 2 \lambda_k^2), \quad \forall k = 1, \dots, K,$$

ce qui permet de fonder des **test d'hypothèses** ou d'établir des **intervalles de confiance** (asymptotiques) concernant les valeurs propres théoriques λ_k (en effet, on montre que les **va** l_k sont asymptotiquement sans corrélation).

(iv) Dans le cas de l'**analyse normée**, on procède comme pour l'analyse centrée. En adoptant (pour simplifier) les mêmes notations, on montre que les vecteurs des composantes principales empiriques $c(k)$ ($k = 1, \dots, K$) définis en (5)' vérifient les propriétés suivantes :

(a) moyenne nulle : $\bar{c}(k) = e_K' c(k) / e_K' e_K = 0$;

(b) variance égale à une valeur propre de V_T : $s(k,k) = \|c(k)\|^2 / e_K' e_K = l_k$;

(c) covariances croisées nulles : $s(k,l) = c(k)' c(l) / e_K' e_K = 0$, si $l \neq k$;

(d) **décomposition de l' « inertie »** de T selon : $\text{tr } C_T = K = \sum_{k=1}^K s(k,k) = \sum_{k=1}^K l_k$;

(e) la covariance empirique entre un vecteur $c(k)$ et une variable u_l (l -ième colonne de U) vaut $z_{lk} l_k$ et leur coefficient de corrélation linéaire vaut $z_{lk} l_k^{1/2}$. Dans ce dernier cas, pour tester si certaines composantes principales sont identiques, ie l'hypothèse d'égalité (ou d'**homogénéité** au second ordre) suivante :

$$(8) \quad H_0 : \lambda_{J+1} = \dots = \lambda_K,$$

on peut utiliser la **statistique** (cf **test de BARTLETT**) :

$$(9) \quad T_N = A(N, K) \cdot \{B(J, K) - C(J, K)\},$$

avec :

$$(e)_1 \quad A(N, K) = N - 6^{-1} (2K + 11) ;$$

$$(e)_2 \quad B(J, K) = (K - J) \text{Log} \{(K - J)^{-1} \sum_{k=J+1}^K l_k\} ;$$

$$(e)_3 \quad C(J, K) = \sum_{k=J+1}^K \text{Log } l_k .$$

T_N vérifie, sous l'hypothèse H_0 , la **propriété asymptotique** suivante :

$$(10) \quad \mathcal{L}(T_N) \xrightarrow{H_0}_{N \rightarrow \infty} \mathcal{X}_d^2 \quad (\text{loi du chi-deux à } d \text{ degrés de liberté}),$$

avec $d = [(1/2)(K - J)(K - J - 1)]$, où $[.]$ est la fonction **partie entière**.

Par suite, une **région critique** (asymptotique) du test est de la forme :

$$(11) \quad w = [T_N \geq q_{1-\alpha}],$$

où $q_{1-\alpha}$ est le **quantile** d'ordre $1 - \alpha$ de la loi χ_d^2 associé au **niveau** $\alpha \in]0, 1[$ du test.

(v) En pratique, l'élément t_{nk} de T représente la **mesure** d'une variable d'**indice** k (eg **caractère statistique**) effectuée sur une **unité statistique**, ie une **observation** (eg relative à un individu) n . On appelle parfois **composante principale** de T tout vecteur propre de $X'X$ (analyse centrée) ou de $U'U$ (analyse normée) associé à une valeur propre non nulle de cette matrice.

Par ailleurs, il existe des **relations de dualité** entre la matrice $X'X$ (resp $U'U$) et la matrice XX' (resp UU') (cf **analyse des données**, **analyse générale des données**).

On peut souvent donner aux propriétés précédentes une **interprétation géométrique**.

Les méthodes d'acp (notamment, l'analyse normée) sont utilisées en **analyse exploratoire des données**, en particulier lorsque :

(a) aucune théorie précise du **phénomène** générant le tableau T n'est connue ;

(b) T est un « grand » tableau (ie $N \gg 0$ ou $K \gg 0$), généralement associé à une **grande base de données** ;

(c) on souhaite projeter (et représenter graphiquement) le **nuage de points** défini par T dans un sous-espace de « faible » dimension, tout en perdant le moins d'**information** possible (arbitrage entre **complexité** et quantité d'**information**) ;

(d) les K variables considérées sont mesurées avec des **unités de mesure** « hétérogènes » (l'analyse normée porte, en effet, sur des nombres sans dimension pr aux unités ou aux **échelles de mesure**, ie les termes de U).

En général, T est un tableau de nombre réels. L'analyse peut, dans certains cas, être appliquée à des observations discrètes (cf **variable de comptage**) ou résultant de **variables qualitatives** (ie eg codées) : ie $T \in M_{NK}(\mathbf{N})$ ou $T \in M_{NK}(\mathbf{Z})$.

Enfin, dans l'étude d'un **modèle de régression**, on remplace parfois, la (N,K) -matrice X des observations des exogènes $\xi_1, \dots, \xi_K$ par ses composantes principales (cf **quasi-colinéarité**, **régression sur composantes principales**).