ANALYSE CANONIQUE (I, K6)
(20 / 12 / 2019, © Monfort, Dicostat2005, 2005-2019)

L'**analyse canonique** constitue une méthode d'**analyse des données** (H. HOTELLING) visant à mesurer l'intensité des liaisons existant entre deux listes de **variables**. Elle constitue une généralisation du **modèle de régression multiple** (linéaire), du **modèle d'analyse de la variance**, de l'**analyse discriminante** ou de l'**analyse factorielle des correspondances**.

(i) Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé**, $\zeta = (\xi, \eta) : \Omega \mapsto \mathbf{R}^K \times \mathbf{R}^G$ un couple de **vecteurs aléatoires** $\xi : \Omega \mapsto \mathbf{R}^K$ et $\eta : \Omega \mapsto \mathbf{R}^G$ (cf **couple aléatoire**). On suppose que ζ est de carré intégrable et l'on pose :

(0) $$V\,\zeta = \begin{pmatrix} \Sigma_{\xi\,\xi} & \Sigma_{\xi\,\eta} \\ \Sigma_{\eta\,\xi} & \Sigma_{\eta\,\eta} \end{pmatrix}$$

avec $\Sigma_{\xi\xi} = V\,\xi$, $\Sigma_{\eta\eta} = V\,\eta$ et $\Sigma_{\xi\eta} = C\,(\xi, \eta) = \Sigma_{\eta\xi}'$ (matrices des **covariances** propres et matrice des **covariances** croisée).

On appelle **analyse canonique** (théorique) de ζ l'étude du **problème d'optimisation** qui consiste à maximiser le **coefficient de corrélation** (linéaire) entre les **formes linéaires** $h'\,\xi$ et $l'\,\eta$ sous des contraintes exprimant que leurs **variances** sont unitaires, ie :

(1) $$\max_{(h,\,l) \in \mathbf{R}K \times \mathbf{R}G} h'\,\Sigma_{\xi\,\eta}\,l$$
$$\text{sous } h'\,\Sigma_{\xi\xi}\,h = 1 \text{ et } l'\,\Sigma_{\eta\eta}\,l = 1,$$

La solution résulte des étapes suivantes :

(a) trouver deux formes linéaires $\alpha_1 = h_1'\,\xi$ et $\beta_1 = l_1'\,\eta$ tq leur coefficient de corrélation linéaire soit maximum : on l'appelle premier **coefficient de corrélation canonique** (théorique) ;

(b) puis, trouver deux formes linéaires $\alpha_2 = h_2'\,\xi$ et $\beta_2 = l_2'\,\eta$ tq $(\alpha_2\,,\,\beta_2)$ soit indépendant de (ie non corrélé avec) $(\alpha_1\,,\,\beta_1)$ et tq leur coefficient de corrélation linéaire soit maximum : on l'appelle deuxième coefficient de corrélation canonique (théorique) ;

(c) etc.

On montre que le problème (1) revient à résoudre les **équations caractéristiques** (théoriques) suivantes :

(2) $$(\Sigma_{\xi\eta}\,\Sigma_{\eta\eta}^{-1}\,\Sigma_{\eta\xi} - \lambda\,.\,\Sigma_{\xi\xi})\,h = 0$$

(3) $(\Sigma_{\eta\xi}\,\Sigma_{\xi\xi}^{-1}\,\Sigma_{\xi\eta} - \mu\,.\,\Sigma_{\eta\eta})\,l = 0.$

Soit $\lambda_1 \geq ... \geq \lambda_K \geq 0$ les **valeurs propres** de (2) et $\mu_1 \geq ... \geq \mu_G \geq 0$ celles de (3). On montre que les valeurs propres non nulles de (2) sont (globalement) les mêmes que celles de (3) ; seuls les ordres de multiplicité diffèrent. Il existe donc une suite $\rho_1 \geq ... \geq \rho_{min(K,G)} \geq 0$ formée des valeurs λ_k (ou μ_g).

On appelle **corrélation canonique** (**théorique**) l'une quelconque des valeurs ρ_1 ,..., ρ_r non nulles (avec $r \leq$ min (K,G)) et **variable canonique** (**théorique**) l'une quelconque des valeurs de la forme linéaire $h'\,\xi$ (resp $l'\,\eta$) correspondant à l'un des vecteurs propres h_1 ,..., h_K (resp l_1 ,..., l_G).

On montre que $r = rg\,\Sigma_{\xi\eta}$.

D'autre part, si l'on procède à l'**analyse en facteurs communs et spécifiques** de ξ et η selon le modèle :

(4)
$$\xi = A\,\varphi + \varepsilon\,,$$
$$\eta = B\,\varphi + \nu,$$

on montre que le plus petit nombre m de facteurs φ communs à ξ et η (ou **nombre effectif** de facteurs communs) compatible avec les hypothèses habituelles :

(5)
$$C(\varphi, \varepsilon) = 0,\ C(\varphi, \nu) = 0,\ C(\varepsilon, \nu) = 0,$$
$$V\,\xi = \Sigma_{\xi\xi} = A\,A' + V\,\varepsilon,$$
$$V\,\eta = \Sigma_{\eta\eta} = B\,B' + V\,\nu,$$
$$C(\xi\,,\,\eta) = \Sigma_{\xi\eta} = A\,B',$$

n'est autre que $m = r = rg\,\Sigma_{\xi\eta}$.

(ii) Dans le cas où G = 1, on obtient le **modèle de régression multiple** linéaire entre η et ξ. Dans le cas général (ie G > 1), la méthode peut poser des problèmes d'interprétation concrète.

(iii) En pratique, si $Z = (Z_1, ..., Z_N)$ est un **échantillon aléatoire** issu de la même **loi** que ζ, on peut estimer ρ_1 ,..., ρ_r à l'aide des solutions r_1 ,..., r_r de l'équation en r :

(6) $$|R_{XY}\,R_{YY}^{-1}\,R_{XY}' - r\,R_{XX}| = 0,$$

qui sont donc les valeurs propres de $R_{XY}\,R_{YY}^{-1}\,R_{XY}'\,R_{XX}^{-1}$, les notations étant tq la matrice des corrélations empiriques de Z s'écrive sous la forme :

(7) $$R_N = \begin{pmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{pmatrix}$$

où R_{XX} est la **matrice des corrélations** empiriques de $X = (X_1, ..., X_N)$, R_{YY} celle de $Y = (Y_1, ..., Y_N)$ et $R_{XY} = R_{YX}'$ celle des coefficients de corrélation empiriques croisés entre X et Y.

De même, les vecteurs a_k ($k = 1, ..., K$) qui estiment les vecteurs h_k ($k = 1, ..., r$) sont les vecteurs propres de la même matrice $R_{XY} R_{YY}^{-1} R_{YX}' R_{XX}^{-1}$, les estimateurs b_g des l_g ($g = 1,...,G$) étant donnés par la relation $b_g = a_g R_{XY} R_{YY}^{-1}$ (en général, on choisit des vecteurs unitaires $||a_k|| = 1$ et $||b_g|| = 1$).

Si l'on suppose que $\rho_1 > ... > \rho_r$ et que $r_1 > ... > r_r$ (inégalités strictes), on peut tester (α étant donné) l'**hypothèse de base** :

(8) $H_0 : \rho_\alpha = 0$

à l'aide de la **statistique** :

(9) $K_N^2 = - \{N - 2^{-1} (K + G + 3)\} \text{Log} \{\prod_{\beta=\alpha}^{r} (1 - r_\beta^2)\}$.

Sous l'hypothèse H_0 , on a :

(10) $\mathcal{L} (K_N^2) \rightarrow_{N \to \infty} \mathcal{X}^2 (K-\alpha+1)(G-\alpha+1)$ (**loi du chi-deux**),

et le test admet pour **région critique** asymptotique, associée au **risque de première espèce** α, la région :

(11) $w = [K_N^2 \geq q_{1-\alpha}]$,

où $q_{1-\alpha}$ est le **quantile** d'ordre $1 - \alpha$ de la loi du chi-deux précédente.

(iv) Du point de vue géométrique, si E_X est le sous-espace (vectoriel) engendré par X dans l'**espace d'observation** $\mathbf{R}^N$ et E_Y celui engendré par Y, il existe une décomposition en **somme directe** :

(12) $E_Y = (E_X \cap E_Y) \oplus E_X^{\perp} \oplus E$,

où $E_X^{\perp}$ est le sous-espace vectoriel orthogonal à E_X et E le sous-espace supplémentaire (dans E_Y) du sous-espace $E_Y \oplus E_X^{\perp}$.

Si $E_X = E_Y$, on dit que X permet de « **prévoir** » Y. Sinon, on dit que les éléments de $E_X \cap E_Y$ sont des **éléments (exactement) prévisibles** à l'aide de ceux de E_X . Les éléments de E_X sont non prévisibles à l'aide de E_X ; enfin, les éléments de E sont des **éléments partiellement prévisibles** à l'aide de E_X .

Dans ce qui précède, un élément est dit « **prévisible** » s'il existe une combinaison linéaire d'éléments de E_X égale à cet élément.