

MATRICE DE CORRÉLATION (C5, F3, J, K, N)

(01 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

Directement fondée sur la notion de **coefficient de corrélation**, celle de **matrice de corrélation** intervient souvent en **analyse multidimensionnelle** : **corrélations croisées**, **analyse des données**, **modèle de régression** ou **modèle d'interdépendance**.

(i) Soit $(\Omega, \mathcal{F}, P)$ un **espace probabilisé** et $\xi : \Omega \mapsto \mathbf{R}^K$ un **vecteur aléatoire** de carré intégrable ($\xi \in \mathcal{L}_{\mathbf{R}^K}^2(\Omega, \mathcal{F}, P)$) dont la **loi de probabilité** est notée P^ξ . On note $\rho_{kl} = \rho(\xi_k, \xi_l)$ le **coefficient de corrélation linéaire** (théorique) entre les coordonnées ξ_k et ξ_l de ξ .

On appelle **matrice de corrélation**, ou **matrice des corrélations**, ou simplement **corrélation, (théorique)** de ξ la **matrice** carrée, notée C_ξ (ou R_ξ , ou encore Γ), dont le terme général est ρ_{kl} .

On note $\sigma = (\sigma_1, \dots, \sigma_K)'$ le vecteur colonne des **écarts-types** des coordonnées de ξ (ie si $\sigma_k^2 = V \xi_k$, $\forall k \in N_K^*$) et $\sigma^\wedge$ la matricialisée de σ (cf **matricialisation**). Alors l'**application** :

$$(1) \quad q : h \in \mathbf{R}^K \mapsto E \{h' (\sigma^\wedge)^{-1} (\xi - E \xi)\}^2 = q(h) \in \mathbf{R}$$

est une **forme quadratique** (semi-)définie positive qui se développe selon :

$$(2) \quad q(h) = \sum_{k=1}^K \sum_{l=1}^K \rho_{kl} h_k h_l = h' (C_\xi) h,$$

où $C_\xi = (\sigma^\wedge)^{-1} E (\xi - E \xi)(\xi - E \xi)' (\sigma^\wedge)^{-1}$ est la matrice de corrélation de ξ , ie la matrice définissant la forme quadratique q .

(ii) On montre que :

(a) C_ξ est reliée à la **matrice de covariance** V_ξ de ξ par la relation :

$$(3) \quad C_\xi = (\sigma^\wedge)^{-1} (V_\xi) (\sigma^\wedge)^{-1},$$

où $\sigma \in D_K(\mathbf{R})$ (**matrice diagonale** d'ordre K) ;

(b) $\text{Diag } C_\xi = \{1, \dots, 1\}$;

(c) $|\rho_{kl}| \leq 1$ (cf **inégalité de CAUCHY-SCHWARZ**) ;

(d) $0 < \text{Dét } C_\xi \leq 1$;

(e) $\sup (\text{Sp } C_\xi) < K$;

(f) $\text{Dét } (V_\xi) = (\prod_{k=1}^K \sigma_k^2) \cdot \text{Dét } (C_\xi)$;

(g) $C_\xi \geq 0$ (**matrice définie positive**).

(iii) Soit $X = (X_1, \dots, X_N)$ un **échantillon iid** issu de ξ , ie un **échantillon indépendant** dont chaque coordonnée $X_n \sim P^\xi, \forall n \in N_N^*$.

On note c_{kl} le coefficient de corrélation linéaire empirique, calculé entre les colonnes x_k et x_l de la matrice $X = (X_1 // X_N) \in M_{NK}(\mathbf{R})$, ie :

$$(4) \quad c_{kl}^2 = x_k' P x_l / \{ \|P x_k\|^2 \cdot \|P x_l\|^2 \}, \quad \forall (k, l) \in (N_K^*)^2,$$

où l'on note de la même façon la matrice et l'échantillon X , et où le symbole $//$ désigne les sauts de lignes reliant les lignes X_1 et X_N de X .

On appelle alors **matrice de corrélation (empirique)** de X (ou de ξ) la (K, K) -matrice dont le terme général est c_{kl} . Cette matrice est notée $C_N X$ ou $R_N X$ (ou encore $C_N \xi$ ou $R_N \xi$).

De façon explicite (grandeur de dimension 0 pr aux **unités de mesure**) :

$$(5) \quad c_{kl}^2 = C_{kl}^2 / (V_k \cdot V_l),$$

où, $\forall (k, l)$:

$$(a) \quad C_{kl} = N^{-1} \sum_{n=1}^N (X_{nk} - \bar{x}_k) (X_{nl} - \bar{x}_l) \text{ (covariance empirique entre } X_k \text{ et } X_l) ;$$

$$(b) \quad V_k = N^{-1} \sum_{n=1}^N (X_{nk} - \bar{x}_k)^2 \text{ (variance empirique propre à } X_k) ;$$

$$(c) \quad \bar{x}_k = N^{-1} \sum_{n=1}^N X_{nk} \text{ (**moyenne empirique** du vecteur colonne } x_k).$$

(iv) Si $\xi \sim P^\xi = \mathcal{N}_K(\mu, \Sigma)$ (**loi normale multidimensionnelle**), on peut tester l'**hypothèse de non corrélation**, parfois appelée **hypothèse d'indépendance**, suivante :

$$(6) \quad H_0 : C \xi = I_K.$$

Le test est fondé sur la **statistique de test** :

$$(7) \quad T_N = -\{N - (1/6)(2K + 1)\} \cdot \text{Log } |C_N X|,$$

avec $|A| = \text{Dét } A$ (cf **déterminant**).

Si l'hypothèse H_0 est vraie, on montre que :

$$(8) \quad \mathcal{L}(T_N) \rightarrow_{N \rightarrow +\infty} \mathcal{X}_d^2$$

(**loi du chi-deux** à $d = C_K^2 = K(K-1)/2$ **degrés de liberté**). On en déduit la **région critique** asymptotique associée à un **risque de première espèce** α :

$$(9) \quad w = [T_N \geq q_{1-\alpha}],$$

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