MATRICE DE COVARIANCE (C5, F3, J, K, N)
(26 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

La notion probabiliste de **matrice de(s) covariance(s)** est fondée sur celle de covariance.

En **Statistique** (**analyse multidimensionnelle**), ce concept joue le rôle de **caractéristique** de **dispersion** pour un vecteur aléatoire ou une **statistique**.

(i) Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé** et $\xi \in \mathcal{L}_{\mathbf{R}K}^2 (\Omega, \mathcal{T}, P)$ un **vecteur aléatoire** de carré intégrable (où **R**K désigne $\mathbf{R}^K$).

On note :

(0) $\quad v_{kl} = C(\xi_k, \xi_l) = E(\xi_k - E\,\xi_k)(\xi_l - E\,\xi_l)$

la **covariance** (théorique) entre une coordonnée ξ_k et une coordonnée ξ_l de ξ, $\forall\ (k, l) \in (N_K^*)^2$.

On appelle alors **matrice de covariance**, ou **matrice des covariances**, ou **matrice des variances-covariances**, ou encore **matrice de dispersion**, ou **matrice du second ordre**, **(théorique)** de ξ la (K,K)-**matrice**, notée $V\,\xi$ (ou $C(\xi, \xi)$, ou $Cov(\xi)$), ou $D\,\xi$, dont le terme général est v_{kl} .

On peut donc aussi la définir selon :

(1) $\quad V\,\xi = E(\xi - E\,\xi)(\xi - E\,\xi)'$.

L'application :

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

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

(3) $\quad q(h) = \Sigma_{k=1}^K \Sigma_{l=1}^K v_{kl}\, h_k\, h_l = h'(V\,\xi)\, h,$

où $V\,\xi$ est la matrice de covariance de ξ, qui définit donc la forme q.

(ii) On montre que :

(a) $(V\,\xi)' = V\,\xi \in S_K(\mathbf{R})$ (**matrice symétrique**) ;

(b) $V\,\xi = E\,\xi\,\xi' - (E\,\xi)(E\,\xi)'$ (**formule de KOENIG-HUYGENS**) ;

(c) $V\,\xi$ est une **matrice de type positif** (au sens de sa forme quadratique associée (3)), ie : $V\,\xi \geq 0 \Leftrightarrow u'(V\,\xi)\,u \geq 0,\ \forall\ u \in \mathbf{R}^K$;

(d) $E\ (u'\ \xi)(v'\ \xi)' = u'\ (E\ \xi\ \xi')\ v,\ \forall\ (u, v) \in (\mathbf{R}^K)^2$;

(e) $C\ (u'\ \xi, v'\ \xi) = u'\ C\ (\xi, \xi)\ v = u'\ (V\ \xi)\ v,\ \forall\ (u, v) \in (\mathbf{R}^K)^2$;

(f) $V(A\ \xi) = A\ (V\ \xi)\ A',\ \forall\ A \in M_{LK}\ (\mathbf{R})$, pour tout $L \geq 1$.

(iii) Soit $X = (X_1, ..., X_N)$ un **échantillon iid** comme ξ (**variable parente**).

On appelle **matrice de(s) covariance(s)**, ou **matrice de dispersion**, **(empirique)** de X (ou de ξ) la (K,K)-matrice, notée S_N ou $S_N\ X$ (ou parfois D_N ou $D_N\ X$), définie par :

(4) $S_N = N^{-1}\ \Sigma_{n=1}^N\ (X_n - \bar{X}_N)\ (X_n - \bar{X}_N)'$,

ie la matrice de dispersion de ξ calculée à l'aide de la **loi empirique** définie par $P_N = N^{-1}\ \Sigma_{n=1}^N\ \delta_{X(n)}$ (où X(n) désigne X_n) (cf **statistique naturelle**). Si l'on définit la matrice aléatoire $X = (X_1' /// X_N')$ (cf **matrice stochastique**), on peut aussi écrire :

(5) $S_N = N^{-1}\ X'\ P\ X = X'\ P\ X / e_N'\ e_N$,

où P désigne la **matrice de centrage par rapport à la moyenne** empirique vectorielle $\bar{X}_N$ et /// des sauts de lignes.

(iv) On établit que :

(a) la matrice $(N / (N - 1))\ S_N$ est un **estimateur sans biais** de $V\ \xi = \Sigma$, ie :

(6) $E\ \{(N - 1)^{-1}\ N\ S_N\} = \Sigma$;

(b) de plus, si $\xi \sim \mathcal{N}_K\ (\mu, \Sigma)$ (**loi normale multidimensionnelle**), alors S_N suit une **loi de WISHART** de paramètres K, N - 1 et $\Sigma / (N - 1)$.

(iv) Comme toute matrice symétrique, une matrice de covariance possède au plus $N\ (N + 1) / 2$ termes indépendants (au sens mathématique courant) (le plus souvent, elle en possède moins).

Ainsi, dans le **modèle de régression multiple** non linéaire (exprimé dans un **espace d'observation**) $y = F\ (b) + u$, où $E\ u = 0$ et $V\ u = \Sigma$, on a $\Sigma \in S_N\ (\mathbf{R})$ (**matrice symétrique** réelle) :

(a) si $\Sigma = \sigma_u^2\ .\ I_N$ (**matrice scalaire**), il existe un seul **paramètre** scalaire : σ_u^2 ;

(b) si Σ est une **matrice diagonale**, il existe N paramètres scalaires : les covariances $\sigma_{\alpha\beta}$ entre **perturbations aléatoires** u_α et u_β ;

(c) si la perturbation u est autocorrélée d'ordre 1 (ie si $u_t = \rho\, u_{t-1} + v_t$, avec $E\, v_t = 0$ et $C\,(v_s\ ,\ v_t) = \sigma_v^2\ \delta_{st}$, $\forall\ (s, t) \in \{1\ ,...,\ T\}^2$), alors Σ admet trois paramètres scalaires : $(\sigma_u^2, \sigma_v^2, \rho)$ (cf **autocorrélation**).

(v) Soit $\xi \in L_{\mathbf{R}K}^2\,(\Omega, \mathcal{T}, P)$ et $\eta \in L_{\mathbf{R}L}^2\,(\Omega, \mathcal{T}, P)$ deux **vecteurs aléatoires**.

On appelle **matrice de(s) covariance(s)** (ou simplement **covariance** ou **dispersion**) de ξ et η la (K,L)-matrice $C\,(\xi, \eta)$ définie selon :

(6) $\quad C\,(\xi, \eta) = E\,(\xi - E\,\xi)(\eta - E\,\eta)' \in M_{KL}\,(\mathbf{R})$.

En particulier, si L = K et $\eta = \xi$, on obtient $C\,(\xi, \xi) = V\,\xi$ (ce qui justifie la notation $C\,(\xi, \xi)$ pour $V\,\xi$).

Dans certains contextes, la matrice $C\,(\xi, \eta)$ définie en (6) est appelée **matrice d'intercovariance** (ou **matrice des inter-covariances**), ou encore **matrice d'interdispersion**, ou encore **matrice de covariance croisée** ou **matrice de dispersion croisée**.

Les mêmes définitions et terminologies valent pour les analogues empiriques.

(vi) On dit que la **loi** P^ξ d'un vecteur aléatoire du second ordre ξ est une **loi de rang r** ssi $rg\,(V\,\xi) = r \leq K$.

P^ξ est appelée **loi régulière** (resp **loi singulière**) ssi r = K (resp $r < K$) (cf **loi dégénérée**, **loi singulière**).