MODÈLE D'INTERDÉPENDANCE (J1)
(25 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

Un **modèle d'interdépendance** est un **modèle statistique** général reliant généralement des **variables quantitatives** entre elles. Parmi ces variables, on distingue des variables endogènes (ou **variables dépendantes**, au sens mathématique du terme) η et des variables exogènes (ou **variables indépendantes**, dans le même sens) ξ.

Ce sont le plus souvent les théories en rapport avec le **phénomène** considéré qui distinguent ces deux types de variables (cf **modèle**), d'où une possibilité d'existence de théories alternatives (théories concurrentes, voire théories « opposées ») (cf aussi **relation fonctionnelle**, **régression**).

(i) On appelle **modèle d'interdépendance**, ou **modèle à équations simutanées**, ou **modèle à équations interdépendantes**, ou **modèle structurel**, ou **modèle à plusieurs équations**, ou encore **modèle d'ensemble**, un modèle défini à partir des données de bases suivante :

(a) un **espace probabilisable** fondamental $(\Omega, \mathcal{T})$ sur lequel est construit un **modèle statistique** fondamental (ou modèle initial) $(\Omega, \mathcal{T}, \mathcal{P})$. Généralement, cet espace et ce modèle ne sont pas explicités (possible **complexité**, ignorance de leur nature précise) ;

(b) deux **espaces d'observation** $(\mathcal{X}, \mathcal{B})$ et $(\mathcal{Y}, \mathcal{C})$. Le premier, $(\mathcal{X}, \mathcal{B})$, est appelé **espace des exogènes**, ou même **espace exogène**. Le second, $(\mathcal{Y}, \mathcal{C})$, est appelé **espace des endogènes** ou **espace endogène** ;

(c) un couple de **variables aléatoires** (cf **couple aléatoire**), généralement multiples :

(1) $(\xi, \eta) : \Omega \mapsto \mathcal{X} \times \mathcal{Y}$.

La première va, ξ, est appelée **variable exogène**, ou **« liste » des variables exogènes** (si $\mathcal{X}$ est considéré comme un espace produit), ou **« vecteur » des variables exogènes** (si $\mathcal{X}$ est un **espace vectoriel** de dimension finie).

La seconde, η, est appelée **variable endogène**, ou **« liste » des variables endogènes** (si $\mathcal{Y}$ est un espace produit), ou **vecteur des variables endogènes** (si $\mathcal{Y}$ est un ev de dimension finie) ;

(d) une **application mesurable** :

(2) $f : \mathcal{X} \times \mathcal{Y} \mapsto \mathcal{Y}$,

parfois appelée **fonction d'interdépendance**, ou **fonction de liaison**, tq le couple (ξ, η) soit lié selon le **système d'équations simultanées**, ou **équation d'interdépendance**, suivant(e) :

(3) $f(\xi, \eta) = \varepsilon,$

dans lequel ε est une **va** définie par (3) elle-même, appelée **variable perturbatrice**, **perturbation aléatoire** ou simplement **perturbation**, ou **« liste » des perturbations**, ou **« vecteur » des perturbations** (selon la nature algébrique de $\mathcal{Y}$).

Autrement dit, $P^{(\xi, \eta)}$ ou $\mathcal{L}(\xi, \eta) = (\xi, \eta)(P)$ est la **loi de probabilité** (ie la **loi conjointe**) de (ξ, η), ie l'image de P par le **couple aléatoire** (ξ, η)), $\forall P \in \mathcal{P}$. La loi de ε est l'image de $P^{(\xi, \eta)}$ par le **changement de va** défini en (3).

L'**équation « implicite »** (3) définit la **forme structurelle** (théorique) du modèle d'interdépendance. Sous certaines hypothèses (eg celles du **théorème de la fonction implicite** si f est continue), on peut passer de (3) à la **forme « explicite »** (cf **régressions multiples**) :

(4) $\eta = g(\xi, \varepsilon),$

qui définit la **forme réduite** du modèle.

Parfois (ie si $\mathcal{Y}$ est un **espace vectoriel**), (4) admet une particularisation importante avec perturbation ε additive, ie tq :

(5) $\eta = h(\xi) + \varepsilon$.

Toutes les équations précédentes sont exprimées dans l'**espace des variables**, et correspondent à la notion courante de **loi scientifique**.

En général, la fonction f (resp g, resp h) n'est pas considérée comme une **fonction aléatoire**, ce qui explique l'introduction d'une va « supplémentaire » ε, considérée comme résumant l'effet de la partie du phénomène considéré qui « échappe » à la description qu'en fait l'**homme de l'art** ou le **statisticien**. Cette va joue le rôle de « **nuance** » (cf **hasard**) pr à l' « **affirmation** » (cf nécessité) traduite par la fonction f (cf **Statistique et hasard**).

(ii) Souvent, la fonction f dépend d'un **paramètre**, soit $\theta \in \Theta$, avec eg $\Theta \in \mathcal{O}(\mathbf{R}^Q)$, interprétable par l'homme de l'art (eg **paramètre d'intérêt**). Par suite, l'équation (3) s'écrit :

(6) $f(\xi, \eta, \theta) = \varepsilon, \qquad \forall \theta \in \Theta,$

l'équation (4) s'écrit :

(7) $\eta = g(\xi, \theta, \varepsilon)$

et l'équation (5) s'écrit :

(8) $\eta = h(\xi, \theta) + \varepsilon,$

Les fonctions précédentes sont définies sur les espaces correspondants et sont généralement supposées connues (ou données). Dans ce cas, θ est souvent appelé **paramètre principal** du modèle, lequel peut aussi comporter des **paramètres secondaires** (cf infra).

(iii) En pratique, la variable exogène ξ comporte K composantes, ie se présente sous la forme :

(a) soit d'un K-uple $(\xi_1, ..., \xi_K)$, lorsque $\mathcal{X}$ est considéré comme un espace produit $\Pi_{k=1}^{K} \mathcal{X}_k$ comportant K espaces facteurs, ou un espace puissance K-ième d'un espace $\mathcal{X}_0$ donné ;

(b) soit d'un vecteur $(\xi_1, ..., \xi_K)'$ comportant K coordonnées, lorsque $\mathcal{X}$ est un espace vectoriel de dimension K.

De même, la variable endogène η comporte G composantes, ie se présente sous la forme :

(a) soit d'un G-uple $(\eta_1, ..., \eta_G)$, lorsque $\mathcal{Y}$ est un espace produit $\Pi_{g=1}^{G} \mathcal{Y}_g$ comportant G espaces facteurs, ou un espace puissance G-ième d'un espace $\mathcal{Y}_0$ donné ;

(b) soit d'un vecteur $(\eta_1, ..., \eta_G)'$ comportant G coordonnées, lorsque $\mathcal{Y}$ est un espace vectoriel de dimension G.

Ainsi, on peut avoir :

(9) $\mathcal{X} = \Pi_{k=1}^{K} \mathcal{X}_k$, $\mathcal{Y} = \Pi_{g=1}^{G} \mathcal{Y}_g$.

La situation la plus fréquente en pratique est celle où $\mathcal{X}_k = \mathcal{X}_0$ (fixé, $\forall$ k $\in$ N_K*) et $\mathcal{Y}_g = \mathcal{Y}_0$ (fixé, $\forall$ g $\in$ N_G*), avec $\mathcal{X}_0$ = **R** et $\mathcal{Y}_0$ = **R**, ie :

(10) $\mathcal{X} = \mathcal{X}_0^K = \mathbf{R}^K$, $\mathcal{Y} = \mathcal{X}_0^G = \mathbf{R}^G$,

ce qui conduit à écrire les équations {(3),(4),(5)} resp selon :

(11) $f(\xi_1, ..., \xi_K, \eta_1, ..., \eta_G) = \varepsilon$;

$$\begin{matrix} & \eta_1 = g_1(\xi_1,\dots,\xi_K,\varepsilon_1,\dots,\varepsilon_G), \\ (12) & \dots \\ & \eta_G = g_G(\xi_1,\dots,\xi_K,\varepsilon_1,\dots,\varepsilon_G); \end{matrix}$$

$$\begin{matrix} & \eta_1 = h_1(\xi_1,\dots,\xi_K) + \varepsilon_1, \\ (13) & \dots \\ & \eta_G = h_G(\xi_1,\dots,\xi_K) + \varepsilon_G, \end{matrix}$$

et les équation {(6),(7),(8)} selon resp :

$$(14) \quad f(\xi_1,\dots,\xi_K,\eta_1,\dots,\eta_G,\theta_1,\dots,\theta_Q) = \varepsilon;$$

$$\begin{matrix} & \eta_1 = g_1(\xi_1,\dots,\xi_K,\theta_1,\dots,\theta_Q,\varepsilon_1,\dots,\varepsilon_G), \\ (15) & \dots \\ & \eta_G = g_G(\xi_1,\dots,\xi_K,\theta_1,\dots,\theta_Q,\varepsilon_1,\dots,\varepsilon_G); \end{matrix}$$

$$\begin{matrix} & \eta_1 = h_1(\xi_1,\dots,\xi_K,\theta_1,\dots,\theta_Q) + \varepsilon_1, \\ (16) & \dots \\ & \eta_G = h_G(\xi_1,\dots,\xi_K,\theta_1,\dots,\theta_Q) + e_G. \end{matrix}$$

(iv) Dans la mise en oeuvre des méthodes d'**inférence statistique**, on suppose d'emblée que $\mathcal{X}$ (resp $\mathcal{Y}$) est un **espace d'observation** (généralement, un **espace d'échantillonnage**). Celui-ci se présente souvent sous la forme d'un espace produit. Dans le cas d'observations en nombre fini N (modèle à « distance finie ») :

$$\begin{matrix} & \mathcal{X} = \Pi_{n=1}^{N}(\Pi_{k=1}^{K}\mathcal{X}_{nk}), \\ (17) & \\ & \mathcal{Y} = \Pi_{n=1}^{N}(\Pi_{g=1}^{G}\mathcal{Y}_{ng}). \end{matrix}$$

En général, les espaces facteurs précédents sont identiques entre eux (puissances cartésiennes), ie $\mathcal{X}_{nk} = \mathcal{X}_k$ ($\forall$ k $\in$ N_K*) et $\mathcal{Y}_{ng} = \mathcal{Y}_g$ ($\forall$ g $\in$ N_G*). D'où $\mathcal{X} = (\Pi_{k=1}^{K}\mathcal{X}_k)^N$ et $\mathcal{Y} = (\Pi_{g=1}^{G}\mathcal{Y}_g)^N$. En général, $\mathcal{X} = (\mathcal{X}_0^K)^N = \mathcal{X}_0^{KN} = \mathbf{R}^{KN}$, ou le plus souvent $\mathcal{X} = M_{NK}(\mathbf{R})$; de même, $\mathcal{Y} = (\mathcal{Y}_0^G)^N = \mathcal{Y}_0^{GN} = \mathbf{R}^{GN}$, ou le plus souvent $\mathcal{Y} = M_{NG}(\mathbf{R})$.

(v) Lorsque $\mathcal{X} = M_{NK}(\mathbf{R})$ (**espace vectoriel** des (N,K)-**matrices** réelles) et $\mathcal{Y} = M_{NG}(\mathbf{R})$ (ev des (N,G)-matrices réelles), les va ξ et η sont resp notées X et Y (**matrice des observations**, en nombre N, des variables exogènes et endogènes). Le modèle d'interdépendance prend alors généralement le nom de **modèle à équations simultanées**. Pour chaque observation constituée d'une ligne X_n de X et d'une ligne Y_n de Y, une équation tq (6) ou (14) s'écrit :

$$(18) \quad f(X_n, Y_n, \theta) = U_n, \forall\, n \in N_N^*,$$

où U_n est la n-ième ligne d'une (N,G)-matrice aléatoire **inobservable** U, à valeurs dans $M_{NG}(\mathbf{R})$, appelée **matrice perturbante**, ou simplement **perturbation de**

l'équation. Le modèle (18) (eg) s'écrit alors, sous forme compacte, dans l'**espace des observations** selon :

(19) $F(X, Y, \theta) = U$, ou encore $F_X(Y, \theta) = U$,

où $F : M_{NK}(\mathbf{R}) \times M_{NG}(\mathbf{R}) \times \Theta \mapsto M_{NG}(\mathbf{R})$ est une fonction dont la « forme analytique » est donnée.

L'exemple type de modèle d'interdépendance est le **modèle d'interdépendance linéaire** dans laquelle F est linéaire pr à (X, Y).

(vi) Les équations (1) ou eg (19) traduisent une **spécification** statistiquement incomplète du modèle d'interdépendance. Elles correspondent seulement à une approche modélisatrice d'un phénomène, dans laquelle on ne considère qu'une **caractéristique légale** de la loi du phénomène : la **fonction d'interdépendance**, qui est un concept de type fonctionnel attaché à cette loi.

Deux compléments, les **hypothèses stochastiques**, sont nécessaires pour mettre en oeuvre l'inférence statistique portant sur f ou sur θ :

(a) d'une part, la spécification du **comportement « moyen »**, ou **comportement « en tendance centrale »**, de ε (resp de U), ie de f (ξ, η) (resp de F (X, Y)). Si une **caractéristique** de **tendance centrale** C (eg C = E pour l'**espérance**, C = S pour le **mode**, etc) peut être définie pour ε, on pose :

(20) $C\,\varepsilon = C\,f(\xi, \eta) = \gamma$,

où γ est la valeur requise (et fixe) pour cette caractéristique. Ainsi, si les espaces sont vectoriels (réels) et si C = E, on souhaite que (cf **modèle de régression**) :

(21) $E\,\varepsilon = E\,f(\xi, \eta) = 0$.

ie que la liaison « moyenne » entre ξ et η, représentée par f, soit nulle (en moyenne, tout ce qui perturbe la liaison entre ξ et η n'exerce qu'un effet nul) ;

(b) d'autre part, la spécification d'un **comportement en « variabilité »**, ou **comportement en « dispersion »**, de ε (resp de U), ie de f (ξ, η) (resp de F (X, Y)), autour de la valeur centrale précédente. Ainsi, l'équation (21) est complétée à l'aide d'une équation tq :

(22) $V\,\varepsilon = V\,f(\xi, \eta) = \Sigma$,

où Σ est la **matrice des covariances** de ε. Si les N va U_n (lignes de U) sont iid (cf **suite iid**), l'équation (22) se transcrit selon :

(23) $V\,U = I_N \otimes \Sigma$.

(vii) Les principaux problèmes liés à l'étude d'un modèle d'interdépendance sont ceux relatifs à l'**estimation** de f (ou de θ si f est donnée), ou aux **tests d'hypothèses**

qui peuvent être émis par la théorie et qui portent sur f (ou sur θ). L'analyse de ce type de modèles constitue une approche générale d'analyse des modèles (cf **classification des modèles**).

Parmi les méthodes usuelles d'estimation du paramètre θ, on distingue :

(a) les **méthodes à information complète** et les **méthodes à information limitée** ;

(b) les **méthodes directes** (ie appliquées à la forme structurelle) et les **méthodes indirectes** (appliquées à la forme réduite). Une méthode indirecte doit permettre de « remonter » aux paramètres d'intérêt (les « coefficients » de la forme structurelle). En effet, ce sont ces **coefficients structurels** qui sont interprétables par l'homme de l'art. Un problème d'**identifiabilité** se pose alors en général.

(viii) Les principales méthodes d'estimation sont :

(a) la **méthode du maximum de vraisemblance**, souvent gaussienne ;

(b) la **méthode des variables instrumentales** et diverses variantes ;

(c) la **méthode des moindres carrés**, et ses variantes ou généralisations.

A titre d'exemple, et selon que le modèle est un modèle identifié ou non, on peut citer la **méthode des moindres carrés indirects**, la **méthode des doubles moindres carrés**, la **méthode des triples moindres carrés**, celle (de la classe k) de THEIL (cf **estimateur de THEIL**), etc.

(ix) Diverses extensions ont été définies, eg :

(a) lorsque $\theta \in B$, **partie bornée** de Θ (supposé être un **espace métrique**) : modèle d'interdépendance à **paramètres bornés** ;

(b) lorsque ξ (resp η) est à valeurs dans une partie bornée de $\mathcal{X}$ (resp de $\mathcal{Y}$) : modèle d'interdépendance à variables (ou observations) bornées (ou limitées) ;

(c) lorsque les observations ne suivent pas un **modèle d'échantillonnage** simple (ie ne sont pas iid) comme ci-dessus, et notamment lorsqu'elles sont auto-corrélées (dans l'espace ou dans le temps) (cf **modèle avec autocorrélation spatiale**, **modèle avec autocorrélation temporelle**). On définit alors eg la notion de **modèle d'interdépendance dynamique** (modèle d'interdépendance autorégressifs, etc). Ceci est notamment le cas lorsque les observations proviennent d'un **processus stochastique** dont on observe des **trajectoires**.

(ix) Le modèle considéré ci-dessus est supposé être un **modèle d'ensemble**, ie décrivant l'ensemble du phénomène étudié. Par distinction, il existe une notion de **modèle partiel**, s'attachant à décrire seulement une « sphère » du phénomène (sous-phénomène).

Une hypothèse fondamentale, implicite dans ces deux types de **modélisation**, concerne le caractère relativement « localisé » de la formalisation (cf **phénomène**). Autrement dit, la loi de probabilité gouvernant le phénomène d'ensemble, ou seulement une sphère plus restreinte, n'est pas, en général, une loi complète : certaines variables influentes, ou « **variables omises** », ou encore « **omissions** », peuvent ne pas être prises en compte (ignorance, impossibilité, simplification ou **parcimonie**, etc) (cf aussi **lacune**).

La loi $\mathscr{L}(\xi, \eta)$ du couple (ξ, η) étant « partielle », l'**inférence statistique** suppose alors implicitement que cette loi est :

(a) soit une **loi marginale** (raisonnement par **marginalisation**). Dans ce cas, l'analyse ne dépend pas des valeurs des variables omises ;

(b) soit une **loi conditionnelle** (cf raisonnement par **conditionnement**). Dans ce cas, l'analyse dépend des valeurs, généralement inconnues ou **inobservable**, de ces omissions.