MODÈLE SEMI-PARAMÉTRIQUE (G2)
(24 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

(i) Soit $(\Omega, \mathcal{T})$ un **espace probabilisable** et $\mathcal{P}$ une **famille** de **mesures de probabilité** définies sur $\mathcal{T}$.

On dit que $\mathcal{P}$ est une **famille semi-paramétrique** ssi elle peut s'écrire sous la forme (cf **modèle paramétrique**) :

(1) $\mathcal{P} = (P_\theta)_{\theta \in \Theta}$,

dans laquelle $\theta = (\theta', \theta'') \in \Theta = \Theta' \times \Theta''$, où Θ' désigne un **espace vectoriel** (réel) de dimension finie Dim $\Theta'' = Q$ et Θ'' un ensemble non identifiable à un espace vectoriel (réel) de dimension finie.

$\mathcal{P}$ est donc une famille partiellement indexable par un nombre fini de **paramètres** $\theta' \in \Theta'$.

On dit alors que le **modèle statistique** $(\Omega, \mathcal{T}, \mathcal{P})$ est un **modèle semi-paramétrique** ssi $\mathcal{P}$ est une famille semi-paramétrique.

On note parfois aussi :

(1)' $\mathcal{P} = (P_{(\theta, \lambda)})_{(\theta, \lambda) \in \Theta \times \Lambda}$,

où $\Theta \times \Lambda$ désigne le produit d'un **espace vectoriel** (réel) Θ de dimension finie Dim $\Theta = Q$ et d'un ensemble Λ non identifiable à un espace vectoriel (réel) de dimension finie. $\mathcal{P}$ est donc partiellement indexable par un nombre fini de paramètres $\theta \in \Theta$. Généralement, $\Theta = \mathbf{R}^Q$.

(ii) La notion de **semi-paramétrage** se transpose aussi au **modèle image** $(\mathcal{X}, \mathcal{B}, \mathcal{P}^\xi)$ issu du précédent par une va $\xi : \Omega \mapsto \mathcal{X}$, ou au le modèle image $(\mathcal{X}, \mathcal{B}, \mathcal{P}^X)$ résultant d'une **statistique** X.

Il vaut donc aussi pour un **modèle d'échantillonnage** $(\mathcal{X}^N, \mathcal{B}^{\otimes N}, \mathcal{P}^X)$ associé à un **échantillon** $X : \Omega \mapsto \mathcal{X}^N$.

On peut considérer que le concept de modèle semi-paramétrique est « intermédiaire » entre celui de « modèle paramétrique » et celui de « modèle paramétré » (ou « modèle indicé ») :

(a) dans un **modèle paramétrique** image $(\mathcal{X}, \mathcal{B}, \mathcal{P}^{\xi})$ associé à une va ξ, la famille $\mathcal{P}^{\xi}$ s'écrit $(P_{\theta}^{\xi})_{\theta \in \Theta}$, avec eg $\Theta \subset \mathbf{R}^{Q}$ et $Q > 1$: une loi quelconque P_{θ}^{ξ} de cette famille est entièrement connue (ou déterminée) dès que θ l'est. Il en va de même pour un modèle paramétrique associé à un N-échantillon X.

Ainsi, lorsque $\xi \sim \mathcal{N}_K(\mu, \Sigma)$ (**loi normale multidimensionnelle**), si le paramètre $(\mu, \Sigma) \in \mathbf{R}^K \times M_K(\mathbf{R})$ est connu (étude des propriétés légales) ou donné (eg **simulation**), cette loi est entièrement déterminée ;

(b) dans un **modèle paramétré** image $(\mathcal{X}, \mathcal{B}, \mathcal{P}^{\xi})$ associé à une va ξ, la famille $\mathcal{P}^{\xi}$ peut toujours s'écrire sous la forme $(P^{\xi})_{P\xi \in \mathcal{P}\xi}$ (en notant resp $P\xi$ et $\mathcal{P}\xi$ pour désigner P^{ξ} et $\mathcal{P}^{\xi}$) : autrement dit, elle est toujours indexable par elle-même, et P^{ξ} joue seulement un rôle d'**indice**. Mais il n'existe pas de dimension $Q > 1$ tq $\Theta \subset \mathbf{R}^{Q}$. Il en va de même pour un modèle paramétré associé à un N-échantillon X.

(iii) Le modèle semi-paramétrique constitue un cadre naturel pour l'analyse statistique lorsque $\mathcal{P}$ ou P^{ξ} (ou P^{X}) ne sont spécifiées que partiellement : eg certains **modèles de régression** ou **modèles d'interdépendance** (cf **régression**).

Ainsi, une **caractéristique légale** γ de $P^{X} \in \mathcal{P}^{X}$ peut, seule, dépendre d'un **paramètre d'intérêt** tq le paramètre $\theta' \in \Theta'$ défini ci-dessus, où $\gamma = g(P^{X})$ et g est une **application caractéristique** (ie le mode opératoire permettant de passer de P^{X} à sa **caractéristique légale** γ).

(iv) A titre d'exemples :

(a) un **modèle de régression multiple** linéaire peut s'écrire sous forme d'un modèle image (dans l'**espace des observations**) :

(2) $\{\mathcal{X} \times \mathcal{Y}, \mathcal{B} \otimes \mathcal{C}, \mathcal{P}^{(X, y)}\}$,

auquel on associe les hypothèses suivantes (**espaces d'observation**, **famille** des **lois de probabilité** et **espérance conditionnelle** relative à la **matrice d'observation** X des exogènes) :

(3)
$$\mathcal{X} = M_{NK}(\mathbf{R}), \qquad \mathcal{Y} = \mathbf{R}^{N}, \qquad (X, y) : \Omega \mapsto \mathcal{X} \times \mathcal{Y},$$
$$\mathcal{P}^{(X, y)} = \{P^{(X, y)} : E(y - X b / X) = 0, \ \forall b \in \mathbf{R}^{K}\},$$

C'est un modèle semi-paramétrique car il concerne un ensemble de **lp** $P^{(X, y)}$ dont seule l'espérance conditionnelle $E\, y / X = X b$ est indexable par un nombre fini de paramètres réels scalaires $b_k \in \mathbf{R}$ ($\forall k \in N_K^*$).

Il existe donc deux « inconnues » : les lois $P^{(X,\ y)}$ (a priori non indexables par un nombre fini de paramètres) et les **coefficients de régression** formant le vecteur $b \in \mathbf{R}^K$.

Ce modèle peut se transformer en un **modèle paramétrique** si l'on complète sa **spécification**, ie si l'on ajoute eg l'hypothèse de **normalité** suivante :

(4) $\mathcal{L}(y / X) = \mathcal{N}_N (X\ b, \Sigma)$ (**loi normale multidimensionnelle**),

dans laquelle $\mathcal{L}$ (y / X) désigne la **loi conditionnelle** de y sachant X : cette loi est astreinte à vérifier E y / X = X b et V y / X = Σ ;

(b) de même, le **modèle d'interdépendance linéaire**, écrit sous la forme d'un modèle image (espace d'observation) :

(5) $\{\mathcal{X} \times \mathcal{Y}, \mathcal{B} \otimes \mathcal{C}, P^{(X,Y)}\}$,

dans lequel :

(6)
$$\mathcal{X} = M_{NK}(\mathbf{R}), \qquad \mathcal{Y} = M_{NG}(\mathbf{R}), \qquad (X, Y) : \Omega \mapsto \mathcal{X} \times \mathcal{Y},$$
$$\mathcal{P}^{(X,Y)} = \{P^{(X,Y)} : E(Y B' + X C' / X) = 0, \qquad \forall (B, C) \in M_G(\mathbf{R}) \times M_{GK}(\mathbf{R})\},$$

est un modèle semi-paramétrique.