MODÈLE À VARIANCE COMPOSÉE (J3, J8, J9)
(14 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

De façon générale, un **modèle à variance composée** est un **modèle** dont le but est d'analyser la décomposition d'une **variable endogène**, en fonction de diverses **variables aléatoires** (eg **variables exogènes**, ou « **facteurs** » d'un **plan d'expérience**) ou de divers **paramètres** aléatoires (cf **relation fonctionnelle**, **modèle de régression**, **fonction de régression**).

Dans le cas (fréquent en pratique) d'un **modèle linéaire**, on étudie notamment la **décomposition de la variance** de l'endogène.

(i) **Premier type de modèle**. Ce cas particulier du **modèle de régression multiple** est directement défini (dans l'**espace des observations**) selon eg la forme non linéaire :

$$(1) \qquad \begin{array}{l} y = F(b) + u, \\ \text{avec } b \in \mathbf{R}^Q,\ E\,u = 0,\ V\,u = \Sigma, \end{array}$$

dans lequel la **matrice de dispersion** Σ de la **perturbation** u (ou de y) est supposée se décomposer selon :

$$(2) \qquad \Sigma = V\,u = V\,y = \sum_{i=1}^{I} \alpha_i \,.\, V_i \,,$$

où les **matrices** $V_i \in S_N(\mathbf{R})$ sont (généralement) supposées connues (en totalité ou en partie) et où $\alpha_i \in \mathbf{R}$, $\forall\, i \in N_I^*$.

Lorsque les matrices V_i sont toutes connues, l'estimation de Σ revient à celle de $\alpha = (\alpha_1, ..., \alpha_I)$.

Pour estimer (b, α), des méthodes usuelles sont :

(a) la **méthode des moindres carrés généralisés**, qui conduit à minimiser la **forme quadratique** $||u||_\Sigma^2 = (y - F(b))'\, \Sigma^{-1}\, (y - F(b))$ pr à b ;

(b) la **méthode du maximum de vraisemblance** ;

(c) une méthode bayésienne (cf **théorie bayésienne**).

En particulier, si (1) est un **modèle linéaire** (avec Q = K et F = X, **matrice d'observation** des exogènes), on utilise souvent l'**estimateur de RAO**.

(ii) **Deuxième type de modèle**. Ce modèle est défini à partir d'un **modèle à erreurs composées**) dans lequel H **facteurs** F_h peuvent influencer une **va** $\eta \in L_{\mathbf{R}}^2(\Omega, \mathcal{T}, P)$.

On note $y_{i(1)\ ...\ i(H),n(i(1)...i(H))}$ l'**observation** de η correspondant à la combinaison $I = (i_1 ,..., i_H)$ des facteurs F_h , où $i_h \in \mathcal{T}_h = \{1,..., I_H\}$, $\forall\ h \in N_H^*$. On note alors $I \in \mathcal{T}$, avec $\mathcal{T} = \Pi_{h=1}^{H} \mathcal{T}_h$, ainsi que $y_{I,n(I)}$ l'observation précédente. Pour alléger la typographie, on note aussi i(h) pour désigner i_h et n(i(1)...i(H)) pour désigner $n_{i(1)...i(H)}$, etc.

Un **modèle à erreurs composées non équilibré, avec intéractions**, est de la forme (toujours dans l'**espace d'observation**) :

$$(3) \qquad y_{I,n(I)} = z + u_{I,n(I)} , \quad \forall\ I \in \mathcal{T},$$

dans laquelle les perturbations $u_{I,n(I)}$ se décomposent selon :

$$(4) \qquad u_{i(1)\ ...\ i(H),n(i(1)...\ i(H))} = (u^{1}_{i(1)} + ... + u^{H}_{i(H)}) + (u^{12}_{i(1),i(2)} + ... + u^{H-1,H}_{i(H-1),i(H)}) + \ ... \ + u^{1...H}_{i(1)...i(H)} + v_{i(1)\ ...\ i(H),n(i(1)...\ i(H))} .$$

Le **modèle à variance composée** associé au modèle {(3),(4)} précédent est alors défini sous les hypothèses suivantes :

(a) $E\ u^{h}_{i(h)} = 0, \qquad \forall\ i_h \in \mathcal{T}_h , \quad \forall\ h \in N_H^* ;$

$E\ u^{hk}_{i(h)i(k)} = 0, \qquad \forall\ (i_h , i_k) \in \mathcal{T}_h \times \mathcal{T}_k , \quad \forall\ (h, k) \in (N_H^*)^2_{<} ;$

... ...

$E\ u^{1...H}_{i(1)...i(H)} = 0, \quad \forall\ I = (i_1 ,..., i_H) \in \mathcal{T} ;$

(b) les va $u^{h}_{i(h)}$, $u^{h,k}_{i(h)i(k)}$ (où $h < k$), ..., $u^{1...H}_{i(1)...i(H)}$ sont indépendantes (ou simplement non corrélées) entre elles ;

(c) $V\ u^{h}_{i(h)} = \sigma_h^2, \qquad \forall\ h \in N_H^* ;$
$V\ u^{hk}_{i(h)i(k)} = \sigma_{hk}^2, \quad \forall\ (h, k) \in (N_H^*)^2_{<} ;$

... ...

$V\ u^{1...H}_{i(1)...i(H)} = \sigma_{1...H}^2 ;$

(d) $V\ v_{In(I)} = \sigma_v^2, \qquad \forall\ I \in \mathcal{T}.$

En pratique, il s'agit d'estimer les **variances** : **variance propre** (ou **variance marginale**) de chaque facteur, ou variance d'une combinaison quelconque de facteurs.

Dans les questions d'**estimation** ou de **test**, les composantes de $u_{l,n(l)}$ définies en (4) sont souvent supposées gaussiennes, leurs **moments** des deux premiers ordres étant ceux définis en {(a),(b),(c)}, ce qui permet de construire la **matrice de covariance** d'ensemble du modèle.

Souvent, la « **constante** » z est elle-même remplacée par un terme variable $z_{l,n(l)}$ qui peut éventuellement dépendre d'un **paramètre** b, ce qui correspond à un **modèle mixte**, ou à un **modèle mixte d'analyse de la variance** : $y_{l,n(l)} = z_{l,n(l)} + u_{l,n(l)} = X_{l,n(l)}\, b + u_{l,n(l)}$.

(iii) **Troisième type de modèle** (parfois appelé **modèle à variance décomposée**, ou **modèle de décomposition de la variance**). Ce modèle est défini (dans l'**espace des variables**) selon :

(5) $\quad \eta = b_0 + \xi'\, b + \varepsilon,$

avec $b_0 \in \mathbf{R}$, $E\ \varepsilon = 0$, $V\ \varepsilon = \sigma_\varepsilon^2$, $E\ b = 0$, $V\ b = \Delta = \text{Diag}\ \{\sigma_1^2\ ,...,\ \sigma_K^2\}$ (**matrice diagonale**), $C\ (\varepsilon, b) = 0$.

Soit (X, y) un N-**échantillon iid** constitué de N **copies** du **couple aléatoire** (ξ, η). L'équation (1) est « observée » selon :

(6) $\quad y = b_0\, .\, e_N + X\, b + u,$

avec $e_N = (1\ ,...,\ 1)' \in \mathbf{R}^N$ (premier vecteur bissecteur), $E\ u = 0$, $V\ u = \sigma_\varepsilon^2\ I_N$, $E\ b = 0$, $V\ b = \Delta$, $C\ (u, b) = 0$. Ce modèle comporte K + 2 paramètres $(b_0\ ,\ \sigma_1^2,...,\ \sigma_K^2,\ \sigma_\varepsilon^2)$. On suppose généralement que $\sigma_l \neq \sigma_k$, $\forall\ (k, l) \in (N_K^*)^2_{\neq}$ (ie les coordonnées de b n'ont pas la même **dispersion** propre).

Le nom donné au modèle (6) vient de l'**équation de décomposition de la variance** suivante (qui résulte des hypothèses) :

(7) $\quad V\ y = X\ \Delta\ X' + \sigma_\varepsilon^2\ I_N\ ,$

dans laquelle $A = X'\ \Delta\ X$ est la **composante** relative à X, et dont le terme général est $a_{\alpha\beta} = \Sigma_{k=1}^{K}\ \sigma_k^2\ x_{\alpha k}\ x_{\beta k}$, et $B = \sigma_\varepsilon^2\ I_N$ est la composante relative à u.

L'équation (7) s'écrit aussi sous une forme comparable à (2) :

(7)' $\quad V\ y = \Sigma_{k=1}^{K}\ \sigma_k^2\ x_k\ x_k'\ + \sigma_\varepsilon^2\ I_N\ ,$

dans laquelle $x_k \in \mathbf{R}^N$ est le k-ième vecteur colonne de X.

Si $\sigma^2 \in \{\sigma_1^2, ..., \sigma_K^2, \sigma_\varepsilon^2\}$, alors, dans la classe $\mathcal{Q}$ des **estimateurs quadratiques** (en y) de σ^2, ie dans :

(8) $\mathcal{Q} = \{T_N : T_N = y' Q y, \forall Q \in M_N^+ (\mathbf{C})\}$,

il existe, $\forall \sigma^2 > 0$, un **estimateur sans biais** $T_N^{\sim} = \sigma_N^2$ de σ^2 ssi les matrices $x_1 x_1'$,..., $x_K x_K'$ et I_N sont linéairement indépendantes. Autrement dit, le vecteur $(\sigma_1^2,..., \sigma_K^2, \sigma_\varepsilon^2)$ est estimable ssi le système constitué par les K + 1 matrices précédentes est un système libre (cf **estimabilité**). Par suite, $E\ T_N^{\sim} = \sigma^2$, $\forall \sigma^2 \in (\sigma_1^2,..., \sigma_K^2, \sigma_\varepsilon^2)$ (cf **estimateur de RAO**).

Pour estimer les variances σ_k^2 ($k \in N_K^*$) et σ_ε^2, on suppose souvent la **normalité** de b (eg pour appliquer la **méthode du maximum de vraisemblance**), ie :

(9) $b \sim \mathcal{N}_K (0, \Delta)$.

Il en va de même en matière de **test**. S'il existe des **indices** $k_\alpha \in N_K^*$, avec $\alpha \in N_L^*$ et $L \leq K$, tq $\sigma_{k(1)}^2 = ... = \sigma_{k(L)}^2$, on peut regrouper les coordonnées correspondantes de b et les colonnes correspondantes de X. La démarche est alors, formellement, analogue à la précédente.

Le modèle (6) se généralise directement, eg selon l'équation :

(10) $y = b_0 e_N + X b + Z d + u$,

dans laquelle $X \in M_{NK}$ (**R**) est une (N,K)-matrice tq la matrice précédente, $Z \in M_{NM}$ (**R**) est une (N,M)-matrice de variables exogènes supplémentaires et $d \in \mathbf{R}^M$ un paramètre vectoriel certain (cf **modèle mixte d'analyse de la variance**).

La forme « équilibrée » (6) contient un cas particulier « non équilibré » dans lequel le **vecteur aléatoire** y (à valeurs dans $\mathbf{R}^N$) se décompose selon :

(11) $y = b_0 . e_N + \Sigma_{l=1}^L X(l) b (l) + u$,

où e_N est le premier vecteur « bissecteur » de $\mathbf{R}^N$ (associé à la « **constante** » b_0 du modèle), les $X(l) \in M_{N,H(l)}$ (**R**) sont des matrices données et certaines, b_0 est un paramètre scalaire certain, les b (l) sont des paramètres vectoriels aléatoires (qui représentent des « **effets** aléatoires ») resp à valeurs dans $\mathbf{R}^{H(l)}$ ($\forall l \in N_L^*$) et u est une perturbation vectorielle aléatoire.

L'analyse statistique de (11) suppose que :

(a) les coordonnées $b_{h(l)} (l)$ (où $h_l = 1,..., H_l$) de b (l) vérifient, $\forall l \in N_L^*$:

(12) $b_{h(l)} (l) \sim \mathcal{N}_1 (0, \sigma^2 (l))$ (**loi normale**)

et sont indépendantes (ou simplement non corrélées) entre elles ;

(b) les **« effets »** b (l) sont indépendants (ou non corrélés) entre eux ;

(c) les coordonnées u_n ($n \in N_N^*$) de u vérifient :

(13) $u_n \sim \mathcal{N}_1 (0, \sigma_\varepsilon^2)$

et sont indépendantes (ou non corrélées) entre elles ;

(d) les vecteurs b (l) sont indépendants de (ou non corrélés avec) u. Par suite, les hypothèses stochastiques s'écrivent :

(d_1) $b (l) \sim \mathcal{N}_{H(l)} (0, \sigma^2 (l) I_{H(l)})$;

(d_2) $C (b (l), b (m)) = E\, b (l)\, b (m)' = 0$ si $m \neq l$;

(d_3) $u \sim \mathcal{N}_N (0, \sigma_\varepsilon^2 I_N)$ (**loi normale** centrée) ;

(d_4) $C (b (l), u) = E\, b (l)\, u' = 0, \ \forall\, l \in N_L^*$.

Le modèle précédent se résume alors à :

(14) $y \sim \mathcal{N}_N (\mu, \Sigma)$,

avec :

(15)
$$\mu = b_0 . e_N ,$$
$$\Sigma = \Sigma_{l=1}^{L} \sigma^2 (l) . X(l) X(l)' + \sigma_\varepsilon^2 I_N .$$

Le modèle {(11),(12),(13)} est ainsi un modèle à variance (dé)composée. Il est souvent utilisé dans l'étude des **plans d'expérience** (cf **plan randomisé**).

Le modèle {(14),(15)} est dit **modèle à variance composée gaussien**. En effet, Σ se décompose, dans (15), en fonction des variances $\sigma^2 (l)$ et de σ_ε^2, variances appelées **composantes de la variance** (ou de la **dispersion**) Σ.

L'**estimation** du modèle {(14),(15)} peut s'effectuer à l'aide de la **méthode du mv** (gaussienne). Les **paramètres d'intérêt** en sont les $\sigma^2 (l)$, $\forall\, l \in N_L^*$, σ_ε^2 et (éventuellement) b_0 .

L'approche bayésienne consduit à doter le vecteur (σ^2 (1) ,..., σ^2 (L), σ_ε^2) d'une **probabilité a priori** qui charge l'**orthant** positif $\mathbf{R}_+^{L+1}$ (eg produit de **lois gamma** ou encore **loi de DIRICHLET**). On maximise alors la probabilité a posteriori pr aux variances en question (cf **probabilité a priori**).