Statistique et analyse des données

Jean-Pierre Marciano

# Sur l'estimation dans les systèmes interdépendants non linéaires

*Statistique et analyse des données*, tome 5, nº 1 (1980), p. 51-75
<http://www.numdam.org/item?id=SAD_1980__5_1_51_0>

Statistique et Analyse des Données
1 - 1980 pp. 51-75



# SUR L'ESTIMATION DANS LES SYSTEMES INTERDEPENDANTS NON LINEAIRES

par Jean-Pierre MARCIANO

Directeur de l'Atelier de Prévision
Faculté d'Economie Appliquée d'Aix- Marseille

De nombreux articles récents ont étudié l'estimation des paramètres d'un modèle, sous forme de système interdépendant linéaire, type de modèle noté ici par la suite $M_{LI}$ ; le présent papier veut étendre ce travail dans le cas plus général de systèmes quelconques, linéaires ou non linéaires, notés $M_{GI}$.

Au delà d'une nécessaire synthèse sur les estimateurs existants, l'auteur étudie leurs propriétés asymptotiques et les expérimente car certains résultats théoriques actuels sont malheureusement impraticables. Quand on plaide l'incompatibilité entre vérité et utilité, la mathématisation n'est plus un détour nécessaire du concret à l'abstrait et le retour au concret est ainsi oublié ; ici au contraire les estimateurs obtenus sont expérimentés.

Un estimateur du maximum de vraisemblance ($\widehat{MV}$) sera défini en 2 , étudié, par une méthode numerique originale, appliqué à un exemple ; il apparaîtra en 3 que dans la classe des estimateurs à distance minimale, asymptotiquement, l'estimateur des doubles moindres carrés est moins efficace que $\widehat{MV}$ à information limitée de même que celui des triples moindres carrés par rapport à $\widehat{MV}$ à pleine information (noté aussi $\widehat{MVPI}$), mais sans le respect de l'hypothèse de normalité les estimateurs $\widehat{MV}$ ne sont plus convergents contrairement au cas linéaire. Enfin, l'auteur proposera en 4 un estimateur qui sous certaines conditions sera à la fois convergent et efficace asymptotiquement.

## 1 - LE MODELE

On définit ici un modèle $M_{GI}$ par un système de K équations avec la $i^{ième}$

sous la forme : $\forall$ i $Y_i = g_i (Y_1, \ldots Y_{i-1}, Y_{i+1}, \ldots Y_K, X_1, \ldots X_S, \theta_1, \ldots \theta_D) + U_i$

Chaque variable expliquée, dans chaque équation, est fonction d'une part de variables expliquées dans d'autres équations les endogènes, et d'autre part, de S variables exogènes, explicatives prédéterminées.

Chaque $i^{\text{ième}}$ équation peut s'écrire sous forme implicite, $f_i(Y, X, \theta) = U_i$ avec $i = 1, 2, \ldots K$

Pour l'estimation, on dispose d'un N - échantillon pour (X, Y),

soit, $Y = (Y_1 \ldots Y_K)$ le vecteur des variables endogènes défini dans $R^K$

soit, $X = (X_1 \ldots X_S)$ celui des variables exogènes défini dans $R^S$

soit, $\theta = (\theta_1 \ldots \theta_D)$ celui des paramètres inconnus à estimer

Dans un modèle non linéaire on n'a pas nécessairement $\theta \in R^{K+S}$ mais on peut avoir $\theta \in R^D$ avec $D > K + S$ . Pour chaque équation, on prendra $\theta \in R^D$ quitte à supposer certaines composantes du vecteur $\theta$ nulles.

Soit, enfin, $U_i$ la variable aléatoire appelée résidu avec $\forall$ i, $E(U_i) = 0$.

On a $\text{Var}(U_i) = \sigma_i^2$ , paramètre inconnu,

$\text{Cov}(U_{in}, U_{in'}) = 0$, $\forall$ n, $\forall$ n' $\neq$ n pour des observations de rang n et n'

Soit $\Omega$ la matrice des covariances de la variable vectorielle U.

Pour la $n^{\text{ième}}$ observation on a dans la $i^{\text{ième}}$ équation :

$$y_{in} = g_i (y_{1,n}, \ldots, y_{i-1,n}, y_{i+1,n}, \ldots y_{K,n}, x_{1,n}, \ldots x_{S,n}, \theta_1, \ldots, \theta_D) + u_{i,n}$$

$f_{in}(y, x, \theta) - u_{in} = 0$ sous forme implicite

*Dans la mesure du possible les variables aléatoires seront ici notées en majuscules et leurs valeurs numériques en minuscules.*

On suppose en outre que les différentes fonctions $f_i$ continues ont par rapport aux variables endogènes des dérivées partielles premières et secondes avec un Jacobien non nul.

## 2 - L'ESTIMATEUR DU MAXIMUM DE VRAISEMBLANCE (noté $\widehat{MV}$)

### 2.1 - définition dans un système non linéaire interdépendant (cas de pleine information)

a) préliminaire

Il s'agit après l'avoir défini de signaler les propriétés asymptotiques suivantes de l'estimateur $\widehat{MV}$, en se plaçant dans le $M_{GI}$ ; il s'agit ici du maximum de vraisemblance à pleine information où l'on étudie en bloc tout le système.

- La preuve de la convergence et de la normalité asymptotique de $\widehat{MV}$ dépend essentiellement de l'hypothèse de normalité du terme résiduel (à l'encontre du cas linéaire).

- $\widehat{MV}$ est asymptotiquement plus efficace que l'estimateur non linéaire des triples moindres carrés (notés ici $\widehat{TMC}$). Cette propriété ne sera étudiée qu'après présentation de l'estimateur $\widehat{TMC}$.

b) première étape : transformation de la fonction de vraisemblance

Si les $u_i$ sont normalement distribués, on peut construire une fonction de vraisemblance pour un échantillon donné de N observations des vecteurs X et Y.

Posons ; $f_n = (f_{1n}, \dots f_{Kn})'$ et $\theta = (\theta_1, \dots \theta_D)'$ alors,

$$V = (2\Pi)^{-\frac{KN}{2}} |\Omega^{-1}|^{N/2} \Pi_n [ J_n \exp \{-\frac{1}{2} f'_n \Omega^{-1} f_n\} ]$$

avec $J = \left[\frac{\partial f_i}{\partial y_j}\right]$ d'éléments notés aussi $J_{ij}$

Alors, (1) $\log V = -\frac{KN}{2} \log 2\Pi + \frac{N}{2} \log |\Omega^{-1}| + \sum_n (\log J_n - \frac{1}{2} f'_n \Omega^{-1} f_n)$

Pour rendre log V maximum, il faut annuler les dérivées par rapport aux éléments $\theta_a$ du vecteur $\theta$ ainsi que celles par rapport aux éléments de $\Omega^{-1}$ notés $\bar{s}_{ij}$ (matrice symétrique ).

$$\frac{\partial \operatorname{Log} V}{\partial \bar{s}_{ij}} = \frac{N}{2} s_{ji} - \frac{1}{2} \sum_n f_{in} f_{jn} = 0$$

Souvent, $\Omega$ sera estimé par itérations avec une valeur initiale fixée.

$$\sum_n f'_n \hat{\Omega}^{-1} f_n = \sum_{ij} (\hat{\bar{s}}_{ij} \sum_n f_{in} f_{jn}) = \sum_{ij} \hat{\bar{s}}_{ij} N \hat{s}_{ji} = NK$$

(1) devient :

$$\text{Log } V = -\frac{KN}{2}\text{ Log } 2\,\Pi + \frac{N}{2}\text{ Log } |\Omega^{-1}| + \sum_n \text{Log } J_n - \frac{KN}{2}$$

$$\text{Soit, Log } V = -\frac{KN}{2}\,|1 + \text{Log } 2\,\Pi| + \frac{N}{2}\text{ Log } |\Omega^{-1}| + \sum_n \text{Log } J_n$$

Appelons la partie non constante de cette somme L,

$$L = \frac{N}{2}\text{ Log } |\Omega^{-1}| + \sum_n \text{Log } J_n$$

La maximisation de la vraisemblance revient à la maximisation de L.

Nous définissons l'estimateur du maximum de vraisemblance comme une racine de $\frac{\partial L}{\partial \theta} = 0$ . Une estimation de $\Omega$ peut être obtenue, à $\theta$ fixé, grâce à l'annulation des dérivées de L ; $\Omega$ peut aussi être choisie arbitrairement au départ, unitaire par exemple, puis révisée par itérations.

c) deuxième étape : dérivation de la vraisemblance pour l'optimisation :

Pour estimer $\theta$, il faut donc maximiser L avec,

$$L = \frac{N}{2}\log|\Omega^{-1}| + \sum_n \log J_n = -\frac{N}{2}\log|\Omega| + \sum_n \log J_n$$

$$(2)\quad \frac{\partial L}{\partial \theta_k} = \sum_i \sum_j \frac{\partial L}{\partial s_{ij}}\frac{\partial s_{ij}}{\partial \theta_k} + \sum_i \sum_j \sum_n \frac{\partial L}{\partial J_{ijn}}\frac{\partial J_{ijn}}{\partial \theta_k}$$

Par différentiation,

$$\frac{\partial s_{ij}}{\partial \theta_k} = \frac{1}{N}\sum_n \left(\frac{\partial f_{in}}{\partial \theta_k} f_{jn} + f_{in}\frac{\partial f_{jn}}{\partial \theta_k}\right) \quad \text{et} \quad \frac{\partial J_{ijn}}{\partial \theta_k} = \frac{\partial^2 f_{in}}{\partial \theta_k\,\partial y_j}$$

(2) devient :

$$(3_a)\quad \frac{\partial L}{\partial \theta_k} = -\frac{1}{2}\sum_i \sum_j \{\bar{s}_{ij} \sum_n \frac{\partial f_{in}}{\partial \theta_k} f_{jn} + f_{in}\frac{\partial f_{jn}}{\partial \theta_k})\} + \sum_i \sum_j \sum_n \bar{J}_{jin}\frac{\partial^2 f_{in}}{\partial \theta_k\,\partial y_j}$$

car il faut rappeler que,

$$\frac{\partial \log|\Omega^{-1}|}{\partial \bar{s}_{ij}} = s_{ji} \quad \text{d'où} \quad \frac{\partial L}{\partial s_{ij}} = -\frac{N}{2}\bar{s}_{ij} \quad \text{et} \quad \frac{\partial L}{\partial J_{ijn}} = \bar{J}_{jin}$$

...

De même, on obtient les dérivées secondes :

$$(3_b)\quad \frac{\partial^2 L}{\partial\theta_a\,\partial\theta_b} = \frac{N}{2}\ \{\sum_i\sum_j\sum_k\sum_n \bar{s}_{ik}\frac{\partial s_{jn}}{\partial\theta_a}\bar{s}_{nj}\frac{\partial s_{ij}}{\partial\theta_b} - \sum_i\sum_j \bar{s}_{ij}\frac{\partial^2 s_{ij}}{\partial\theta_a\,\partial\theta_b}\}$$

$$- \sum_i\sum_j\sum_k\sum_n\sum_l \bar{J}_{jkl}\frac{\partial J_{knl}}{\partial\theta_a}\bar{J}_{nil}\frac{\partial J_{ijl}}{\partial\theta_b} + \sum\sum_{ijl} \bar{J}_{jil}\frac{\partial^2 J_{ijl}}{\partial\theta_a\,\partial\theta_b}$$

avec,

$$(4'a)\quad \frac{\partial^2 s_{ij}}{\partial\theta_a\,\partial\theta_b} = \frac{1}{N}\sum_l \{\frac{\partial^2 f_{il}}{\partial\theta_a\,\partial\theta_b} f_{kl} + f_{il}\frac{\partial^2 f_{jl}}{\partial\theta_a\,\partial\theta_b} + \frac{\partial f_{il}}{\partial\theta_a}\frac{\partial f_{jl}}{\partial\theta_b} + \frac{\partial f_{il}}{\partial\theta_b}\frac{\partial f_{jl}}{\partial\theta_a}\}$$

et,

$$(4'b)\quad \frac{\partial^2 J_{ijl}}{\partial\theta_a\partial\theta_b} = \frac{\partial^3 f_{il}}{\partial\theta_a\,\partial\theta_b\,\partial y_j}$$

On montrera en 2.3 qu'une des racines de $\frac{\partial L}{\partial\theta} = 0$ est convergente et que nous avons,

$$\sqrt{N}\ (\hat{\theta} - \theta_0) \rightarrow \mathcal{N}(0,\ - \operatorname{plim}\left[\frac{1}{N}\ \frac{\partial^2 L}{\partial\theta_k\ \partial\theta_{k'}}\right]^{-1})$$

d) troisième étape : Maximisation de la Vraisemblance

Pour maximiser $L$ on peut utiliser la méthode du Gradient et celle de NEWTON-RAPHSON.

Les méthodes de gradient pour maximiser $L(\theta)$ dépendent de l'ajustement d'un paraboloïde à une surface $0 = L(\theta)$ obtenue en considérant le développement de Taylor jusqu'au second ordre de la fonction de vraisemblance prise en un point $\theta°$ dans l'espace des paramètres.

$$0 = L\ (\theta°) + \overrightarrow{\theta°\theta}\ \ (\text{Grad } L)_\circ + \sum_k\sum_{k'} (\theta_k - \theta_k^\circ)\ (\theta_{k'} - \theta_{k'}^\circ)'\ (\frac{\partial^2 L}{\partial\theta_k\ \partial\theta_{k'}})_\circ \quad (5)$$

On choisit un vecteur $V$ et une direction $\lambda V$ dans cet espace, alors,

$$L\ (\theta° + \lambda V) = L\ (\theta°) + \lambda V\ (\text{Grad } L)_\circ + \frac{1}{2}\lambda^2 \sum\sum_{kk'} V_k\ V_{k'}^T\ (\frac{\partial^2 L}{\partial\theta_k\ \partial\theta_{k'}})_o \qquad (6)$$

*La notation $V^T$ indique le transposé d'un vecteur $V$. Cette notation sera préférée au prime quand il y aura risque de confusion.*

Posons $L(\theta°) = L_o$

$(\text{Grad } L)_o = P$

$$- \frac{\partial^2 L}{(\partial\theta_k \, \partial\theta_{k'})_o} = Q_{kk'}, \qquad \text{élément d'une matrice } Q$$

$$\text{alors, } \Theta = L° + \lambda V^T_. P - \frac{1}{2} \lambda^2 V^T Q V \qquad (7)$$

O est maximum si, $V^T_. P - \lambda V^T Q V = 0$

Cette condition est suffisante si $Q$ est définie positive, alors,

$$\lambda = + \frac{V^T_. P}{V^T Q V}$$

(7) devient,

$$L(\theta) = O = L_o + \frac{(V^T P)^2}{V^T Q V} - \frac{1}{2} \frac{(V^T P)^2}{V^T Q V}$$

d'où,

$$\Delta L = L - L_o = \frac{1}{2} \frac{(V^T P)^2}{V^T Q V}$$

quantité positive si $Q$ est définie positive.

Pour choisir V, on essaie de maximiser la partie du 1° ordre de $\theta$ pour une longueur de $V$ fixée définie par une matrice $G$ définie positive.

$$\begin{cases} V^T . P & \text{à maximiser} \\ \text{avec } V^T G V = k^2 \end{cases}$$

la résolution donne,

$$\vec{V} = \frac{k \, G^{-1} P}{(P^T G^{-1} P)^{1/2}}$$

sans perte de généralité, on peut poser $k^2 = P^T G^{-1} P$, alors, (8) $V = G^{-1} P$ et (9) $\lambda = (P^T G^{-1} P) / (P^T G^{-1} Q G^{-1} P)$

Le choix $G = Q$ donne le gain maximum $\Delta L$ pour $(\text{Grad } \vec{L})_o$ donné (<u>méthode de Newton</u>) ; alors $\lambda = 1$

En résumé de cette étape, on veut L maximum, on se donne une valeur initiale de $\theta : \theta_o$ ; (8) nous donne un vecteur $(\vec{V})$ tandis que la deuxième étape nous a permis d'avoir le calcul des dérivées partielles de la vraisemblance. Le maximum de L est obtenu par la formule itérative.

$$\theta = \theta_o + V$$

...

Dans la pratique, on choisit généralement la valeur initiale $\theta_o$ comme une des plus probables subjectivement. Cette approche est voisine de l'approche bayésienne, où il existe une information à priori sur $\theta$ mais qui prend alors la forme d'une distribution de probabilité. Il existe de nombreuses méthodes dérivées de la méthode de Newton et dites de Newton modifiées.

En supposant $L$ concave, nous proposons une méthode d'estimation de $\theta$ (maximisation de L) plus originale et plus appropriée, appelée méthode du gradient conjugué.

Soit donné initialement $\theta_{(o)} \in R^D$ on pose $r_o = - L'(\theta_{(o)})$ et $w_o = r_o$ puis pour $n = 0, 1, \ldots$ définissons,

$$\theta_{(n+1)} = \theta_{(n)} + \mu_n \; w_n$$

Dans le cas quadratique, on prendra $\mu_n = \dfrac{(w_n, r_n)}{(L''.w_n, w_n)}$ $\theta_{(n+1)}$ n'est autre que le point qui maximise $L(\theta)$ sur la droite passant en $\theta_{(n)}$ et de direction $w_n$. Dans le cas non linéaire général $\mu_n$ n'est plus donné par une formule mais est déterminé par une méthode d'approximation quelconque de l'optimum sur la direction $w_n$, méthode de dichotomie par exemple, bien connue en analyse numérique. Puis on pose :

$$r_{n+1} = - L'(\theta_{(n+1)})$$

et,

$$w_{n+1} = r_{n+1} + \lambda_{n+1} \; w_n$$

avec,

$$\lambda_{n+1} = - \frac{(r_{n+1}, L''w_n)}{(r_n, L''w_n)} = \frac{||r_{n+1}||^2}{||r_n \;||^2}$$

$\theta_{(n+1)}$ est cherché à partir de $\theta_{(n)}$ dans une direction particulière $w_n$ dite "de gradient conjugué". De façon générale, M vecteurs ($M \leq D$) $v_1, \ldots v_M \in \mathbb{R}^D$ sont dits A-conjugués par rapport à une matrice symétrique définie positive A si :

$$(Av_i, v_j) = \begin{cases} 0 \text{ si } i \neq j \\ \neq 0 \text{ si } i = j \end{cases}$$

Dans $\mathbb{R}^D$, la méthode converge en moins de $D$ itérations. Il est toutefois prudent de vérifier que $w_n$ est une bonne direction de maximisation, c'est à dire que $(w_n, L'(\theta_{(n)}) > 0$ ; si ce n'est pas le cas, on prend $\theta_{(n)}$ comme nouveau point initial avec $w_n = - L'(\theta_{(n)})$.

Après un cycle de D directions conjuguées, on peut prendre $\theta_{(D)}$ comme nouveau point initial pour un nouveau cycle de gradients conjugués.

2.2 - application économique

Une illustration originale pour la France va être présentée ici par l'auteur.

Soit une fonction de production de type Cobb Douglas, classique en économie, à rendement constant,

$$P = \alpha_1 \, 10^{\alpha_2 X_1} \, L^{\alpha_3} \, K^{1-\alpha_3}$$

où, P est le produit national brut

$X_1$ est le temps ($X_1 = 0, \ldots, 13$) à partir de l'année 1959 jusqu'à 1972

L le travail exprimé par la population active intérieure occupée,

K est le stock de capital (évalué selon la méthode de Mairesse).

On peut, à partir de cette équation, en divisant de part et d'autre par L, exprimer le produit par tête de travailleur ($\frac{P}{L} = Y_1$) comme une fonction du capital par tête, ($\frac{K}{L} = Y_2$) intensité capitalistique de la production sous la forme :

$$(10) \quad Y_1 = \alpha_1 \, 10^{\alpha_2 X_1} \, Y_2^{1-\alpha_3}$$

On peut estimer les paramètres de l'équation (10) en introduisant une variable résiduelle multiplicative et en opérant sur l'équation une transformation logarithmique qui la linéarise. On peut aussi introduire une variable résiduelle additive $U_1$ suivant les hypothèses statistiques classiques décrites plus haut ; on a alors un modèle non linéaire. On remarque, en outre, que le paramètre $\alpha_3$ à estimer représente l'élasticité partielle du facteur travail, son degré de substituabilité au capital. De plus, les fonctions de Cobb-Douglas font l'hypothèse d'une substituabilité des facteurs ; cette hypothèse s'exprimera par une deuxième équation, interdépendante avec la première ; le capital dont dispose en moyenne pour la production chaque travailleur, $Y_2$, est ainsi fonction du prix du facteur travail, du taux de salaire réel $X_2$, le pouvoir d'achat sera le rapport du taux de salaire au prix à la production d'où la deuxième équation,

$$Y_2 = \frac{X_2}{\alpha_3} + U_2$$

...

KLEIN, aux Etats Unis, avait déjà estimé les paramètres d'un modèle similaire avec une série d'observations sur la première moitié au 20e siècle. Ici la série des observations de l'auteur pour les deux pays va de 1959 à 1972. On peut dresser un tableau comparatif de nos résultats pour la France et pour les U.S.A. obtenus équation par équation, par $\widehat{M.C.O}$. puis globalement par $\widehat{MV}$.

| Estimation $\widehat{MCO}$ | | Modèle U.S.A. | | Idem France |
|---|---|---|---|---|
| | | (1909-49 Klein) | (59-72) | (59-72) |
| 1e équation | $\hat{\alpha}_1$ | 0,975 | 1,544 | 1,079 |
| | $\hat{\alpha}_2$ | 0,0081 | 0,0086 | 0,013 |
| | $\hat{\alpha}_3$ | 1,102 | 0,319 | 0,420 |
| 2e équation | $\hat{\alpha}_3$ | 0,660 | 0,500 | 0,362 |

| Estimation $\widehat{MV}$ | | | | |
|---|---|---|---|---|
| | $\hat{\alpha}_1$ | 0,638 | 2,420 | 0,914 |
| | | (0,007) | (0,037) | (0,122) |
| | $\hat{\alpha}_2$ | 0,0074 | 0,011 | 0,0072 |
| | | (0,0003) | (0,0003) | (0,0023) |
| | $\hat{\alpha}_3$ | 0,659 | 0,511 | 0,388 |
| | | (0,002) | (0,004) | (0,073) |

Les estimations de $\widehat{MV}$ ont été obtenues par itérations successives avec pour valeurs initiales celles de l'estimation par $\widehat{M.C.O}$.

On remarque sur la première période que l'estimation par M.C.O. (incorrecte dans ce système interdépendant) donne des résultats incohérents sur $\hat{\alpha}_3$ (> 1) ; de plus, l'estimation de $\alpha_3$ est très différente de la première à la deuxième équation, par contre l'estimation par MV rectifie cette incohérence. Pour cette dernière estimation, les écarts types estimés sont dans l'ensemble très bons pour des coefficients estimés différents de ceux obtenus par M.C.O., de près de 100 % parfois.

Dans un autre domaine, les théories économiques récentes rendent les fonctions de demande de biens durables interdépendantes aussi bien avec les demandes de loisirs qu'avec les systèmes exprimant le niveau de l'emploi. A revenu constant, les ménages peuvent préférer substituer à une demande de biens durables, une demande de loisirs, surtout si le temps devient un bien rare. Cette substituabilité (du même type que celle entre capital et travail dans les fonctions de production) peut dépendre notamment du temps que le bien durable permet de gagner, (machine à laver, ...) du niveau de l'emploi (une mère de famille au chômage a moins besoin d'un four programmable) et de la durée hebdomadaire de travail.

Un système de fonctions de demande de biens incluant la demande de loisirs et la condition de substituabilité n'est généralement pas linéaire, au niveau des paramètres à estimer. La théorie d'estimation des $M_{GI}$ ne permettait pas jusqu'ici une estimation par $\widehat{MV}$ d'un tel système en présence d'échantillons de grande taille.

On va donc chercher dans le $M_{GI}$ un estimateur $\widehat{MV}$ qui, sous certaines conditions de régularité, soit asymptotiquement,

- convergent,
- efficace,
- normal

## 2.3 - les propriétés asymptotiques de $\widehat{MV}$

*1. en présence d'équations indépendantes (système $M_{GN}$)*

Malinvaud dans [9] les a étudiées pour aboutir aux théorèmes 1 et 2 ci-dessous par élargissement du problème suivant :

$(P_1)$ : soit $Y$ un vecteur aléatoire avec $\Omega$ matrice des covariances et, $E(Y) = z$ ($z \subseteq G$, variété linéaire).

Connaissant $\Omega$ et $G$, ayant une observation $y$ de $Y$, on veut estimer $z$ ; le problème $(P_1)$ peut s'élargir à cet autre problème :

$(P_2)$ idem, mais $G$ est un ensemble quelconque ; ce problème est plus général mais ne se prête pas à une théorie générale.

On y définit néanmoins un estimateur à distance minimale $\hat{z}$ qui minimise, [illegible] existe, $|Y - z|$

Soit alors une matrice $S_N$ définie positive, éventuellement aléatoire, utilisée à définir la distance au niveau de l'échantillon.

Soit l'expression pour notre modèle de base, mais avec équations indépendantes.

$$L_N(S_N, \theta) = \sum_n (y_n - g_n(\theta))' \, S_N (y_n - g_n(\theta))$$

et $\hat{\theta}$ minimisant $L_N$ (on suppose $L_N$ et ses 2 premières dérivées continues).

Hypothèses préliminaires aux théorèmes 1 et 2 :

Les hypothèses classiques sur les $U_i$ et les $g_i$ sont conservées. ($i_o$ : les $u_i$ sont notamment indépendantes et de même distribution $(0, \Omega)$) ; viennent s'y ajouter les hypothèses suivantes :

($i_1$) la matrice $S_N$ tend en probabilité vers une matrice régulière

($i_2$) si à priori $\theta \in \emptyset$ $(\emptyset \subseteq \mathbb{R}^D)$

Alors, $\hat{\theta}$ minimise, dans $\emptyset$, $L_N$ et ici $\emptyset$ est supposé borné uniformément, $\theta_o$ étant la vraie valeur de $\theta$.

($i_3$) les variables exogènes sont aussi supposées appartenir à un ensemble X borné de $\mathbb{R}^S$. Leur distribution est stable dans $\Xi$ (leurs valeurs se répartissent comme si elles résultaient de N tirages d'une même loi) et est écarté le phénomène correspondant à la colinéarité.

($i_4$) dans $\mathbb{R}^D$ le domaine $\emptyset$ des valeurs a priori possibles de $\theta$ contient un voisinage V du vecteur $\theta_o$ des vraies valeurs, dans ce voisinage les fonctions $f_i$ sont bornées dans leur ensemble ainsi que leurs dérivées des trois premiers ordres, qui existent, avec,

$$\left\{\begin{array}{l} N \to \infty \\ \quad\Downarrow \\ M_N(\Sigma) \to M(\Sigma) \text{ régulière} \end{array}\right.$$

$\forall\ \Sigma$ matrice définie positive symétrique avec,

$M_N(\Sigma) = \frac{1}{N} \sum_n (P'_n \Sigma P_n)$ où P est la matrice des dérivées premières des fonctions $f_i$ par rapport aux paramètres (le symbole $\sum_n$ est évidemment ici pour sa part le signe somme).

Théorème 1 :

Sous les hypothèses précédentes si $\hat{\theta}$ est convergent, si $S_N$ aléatoire tend vers $\Sigma$ définie positive, alors

$$\sqrt{N}\,(\hat{\theta} - \theta_o)$$

a une loi limite normale quand $N \to \infty$ avec une matrice des covariances calculable.

conséquence : Dans la classe des estimateurs à distance minimale, ceux qui ont la meilleure efficacité asymptotique sont obtenus avec une matrice $S_N$ qui tend en probabilité vers $\Omega^{-1}$.

Théorème 2 :

Même résultat que le théorème 1, mais $\hat{\theta}$ obtenu à partir d'une matrice de covariances observée, on pourra pour la construire partir de $S_o$ quelconque, unitaire par exemple, puis en obtenant $\hat{\Omega}$ par valeurs observées et itérations successives ; cet estimateur est alors celui des moindres carrés généralisés itérés ($\widehat{MCGI}$).

Si le nombre d'itérations devient infini, on ne peut plus affirmer directement que $\hat{\theta}$ est convergent. L'auteur propose la démonstration restrictive suivante :

lemme 1 - soit $\hat{\theta}$ estimateur convergent pour $\theta_o$ vraie valeur de $\theta$. Alors,

$$T(\hat{\theta}) = \Omega_o + o_p\ (1) \quad \text{avec } T(\hat{\theta}) = \frac{1}{N} \sum_n (y_n - g_n(\hat{\theta}))'(y_n - g_n(\hat{\theta}))$$

avec $\Omega_o$ vraie valeur de $\Omega$.

et, $o_p(1)$ indique un développement limité au 1er terme qui tend vers zéro en probabilité.

*On trouve dans le chapitre 9 de MALINVAUD [9] une méthode pour cette démonstration.*

lemme 2 - soit $\hat{\Omega}_N$ suite d'estimations, matrices aléatoires définies positives, A une matrice définie positive et $\hat{\Omega}_N = A + o_p(1)$, $\theta(\hat{\Omega}_N)$ l'estimateur à distance minimale de $\theta_o$ pour $\hat{\Omega}_N$ donnée. Alors, $\theta(\hat{\Omega}_N) = \theta_o + o_p\ (1)$ quand $N \to \infty$

*L'estimateur $\widehat{MCG}$ n'est pas là itéré à partir de $\hat{\Omega}$. Ce résultat se démontre à partir de l'hypothèse (1) et des conditions I à IV dans MALINVAUD [9, ch.9]*

lemme 3 - soit $X_{mn}$ une double suite de v.a.

$X_n$ une suite de v.a., c un scalaire

si $X_{m\,n} \to X_n$ uniformément en n quand $m \to \infty$

si $X_{m\,n} = c + o_p\ (1)$ quand $n \to \infty$ $\qquad \forall\ m$

alors, $X_n = c + o_p\ (1)$

*On va étudier l'estimateur $\widehat{MCG}$ itéré quand m, le nombre d'itérations, tend vers l'infini ainsi que n, nombre d'observations. Pour démontrer le lemme 3, on a*

$$|X_n - c| \leq |X_n - X_{mn}| + |X_{mn} - c|$$

d'où,

$$[Pr\ |X_n - c| > \varepsilon] \leq Pr\ (\ |X_{mn} - c| > \varepsilon - [X_n - X_{mn}]\ )$$

*comme* $X_{mn} \to X_n$ $\exists\ m_o \leq m$ *tel que* $|X_{mn} - X_n| < \delta \quad \forall\ \delta > 0$

*et* $\varepsilon - [X_{mn} - X_n] > \varepsilon - \delta$ *ceci uniformément, d'où l'inégalité en probabilité ci-dessus et la limite de* $X_n$.

Théorème 2' (convergence) :

Sous les hypothèses de régularité du modèle, l'estimateur $\widehat{MCG}$ itéré converge vers une solution $\hat{\theta}_N^+$ de $\frac{\partial V}{\partial \theta +} = 0$, où $V$ est la fonction de vraisemblance et $\hat{\theta}_N^+ = \theta_o^+ + o_p\ (1)$ où l'indice supérieur $+$ indique un ensemble de paramètres incluant $\theta$ et les éléments de $\Omega$.

*La démonstration est une application directe du lemme 3, si l'on suppose que,*

$$\theta_{mN} \to \hat{\theta}_N \text{ uniformément en } N \text{ quand } m \to \infty$$

*avec* $\hat{\theta}_N$ *solution de* $\frac{\partial L}{\partial \theta} = o$ *alors,* $\theta_{mN} \to \theta_o$ *en probabilité quand* $N \to \infty$ *(Lemme 3) et*

$\hat{\theta}_N = \theta_o + o_p\ (1)$

Dans le paragraphe suivant, on doit étudier la convergence de $\hat{\theta}$ qui dépend lui-même du nombre d'itérations Sur les conseils de W. BARNETT [0] nous avons introduit ci-dessous les lemmes conduisant au Théorème 2' pour vaincre la difficulté de la double convergence dans le théorème 3 suivant.

2. *en présence d'équations interdépendantes,*

Il est possible d'étendre ces résultats aux systèmes interdépendants et d'y obtenir un théorème généralisant les théorèmes 1 et 2 de Malinvaud, et le théorème (2') de l'auteur.

Théorème 3 :

Un estimateur $\widehat{MV}$ dans un modèle $M_{GI}$ est asymptotiquement Normal et en outre convergent, c'est à dire qu'une racine $\hat{\theta}$ de l'équation $\frac{\partial L}{\partial \theta}$ est convergente et satisfait,

$$\sqrt{N}\ (\hat{\theta} - \theta_o) \to \mathcal{N}\left(0,\ - \lim N \left(\left[\frac{\partial^2 L}{\partial \theta \partial \theta'}\right]_{\theta_o}\right)^{-1}\right)$$

Mais la démonstration dépend essentiellement de la Normalité de U. Il n'y a pas convergence si la vraie distribution n'est plus celle de l'hypothèse.

Démonstration :

Si au voisinage de $\theta_o$ on développe en série de TAYLOR $\frac{L_N}{N}$ on obtient :

$$\frac{L_N(\theta)}{N} = \frac{L_N(\theta_o)}{N} + \frac{1}{N}\left(\frac{\partial L_N}{\partial \theta'}\right)(\theta - \theta_o) + \frac{1}{2}(\theta - \theta_o)^T \frac{1}{N}\left(\frac{\partial^2 L_N}{\partial\theta\partial\theta'}\right)_{\theta^*}(\theta - \theta_o)$$

avec $\theta^*$ entre $\theta$ et $\theta_o$

Avant d'étudier la limite en probabilité de ce développement, on introduit les hypothèses complémentaires suivantes :

$j_1$) au voisinage $V_o$ de $\theta_o$, $\exists \lim \frac{L_N}{N}$

$j_2$) $\frac{1}{\sqrt{N}}\left(\frac{\partial L_N}{\partial \theta}\right)_o \to \mathcal{N}\left[0, \lim \frac{1}{N} E\left(\left(\frac{\partial L_N}{\partial \theta}\right)_o, \left(\frac{\partial L_N}{\partial \theta'}\right)_o\right)\right]$

$j_3$) en $V_o$, $\exists \lim \frac{\partial^2 L_N}{\partial\theta\partial\theta'}$ avec convergence uniforme

alors,

$$\lim \frac{L_N(\theta)}{N} = \lim \frac{L_N(\theta_o)}{N} + 0 + \frac{1}{2}(\theta - \theta_o)^T \lim \frac{1}{2}\left(\frac{\partial^2 L_N}{\partial\theta\partial\theta'}\right)_{\theta^*}(\theta - \theta_o)$$

On peut introduire une nouvelle hypothèse :

$j_4$) $\lim \left(\frac{\partial^2 L_N}{\partial\theta\partial\theta'}\right)_o < 0$ (si $L_N$ est optimale en $\theta_o$, cet optimum est un maximum)

Alors, il apparaît au vu du développement de TAYLOR, que :

$\lim \frac{L_N(\theta)}{N} < \lim \frac{L_N(\theta_o)}{N}$ ; on a un maximum local à la limite en $\theta_o$ ; donc une racine $\hat{\theta}$ de l'équation $\frac{\partial L}{\partial \theta}$ est convergente grâce aux lemmes (1), (2), (3); le fait que $\hat{\theta}$ soit lui-même fonction du nombre d'itérations, n'est plus un obstacle.

Pour démontrer la normalité asymptotique il suffit d'ajouter une dernière hypothèse, sur les limites en probabilité ,

$$j_5)\quad \lim \frac{1}{N} E\left|\left(\frac{\partial L_N}{\partial \theta}\right)_o \left(\frac{\partial L_N}{\partial \theta'}\right)_o\right| = -\lim \frac{1}{N}\left(\frac{\partial^2 L_N}{\partial\theta\partial\theta'}\right)_o$$

et de remarquer la transformation du développement en série de TAYLOR, remplaçant $\theta$ par $\hat{\theta}$

$$\left(\frac{\partial L_N}{\partial \theta}\right)_{\hat{\theta}} = \left(\frac{\partial L_N}{\partial \theta}\right)_o + \left(\frac{\partial^2 L_N}{\partial\theta\partial\theta'}\right)_{\theta^*} (\hat{\theta} - \theta_o)$$

## 3 - ESTIMATEURS DES DOUBLES ET TRIPLES MOINDRES CARRES ($\widehat{DMC}$, $\widehat{TMC}$)

Ces estimateurs à distance minimale, à information limitée ($\widehat{DMC}$) ou totale ($\widehat{TMC}$), n'apportent jamais la meilleure efficacité asymptotique, mais la preuve de leur convergence est indépendante de la distribution de U.

### 3.1 - l'estimateur $\widehat{DMC}$

Il se déduit d'une méthode à information limitée où on traite le système équation par équation. La définition donnée en 2.3. d'un estimateur à distance minimale, consiste à trouver $\hat{\theta}$ minimisant :

$$L_N (S_N, \theta) = \sum_n (y_n - g_n(\theta))' S_N(y_n - g_n(\theta))$$

Soit pour la $i^{\text{ième}}$ équation du $M_{GI}$

$$L_N (S_N, \theta) = f_i^T S_N f_i \quad \text{avec } f_i = y_i - g_i$$

$$y_i = (y_{i1}, \ldots, y_{iN})$$

$$g_i = (g_{i1}, \ldots, g_{iN})$$

et le modèle initial $Y_i = g_i (Y,X,\theta) + U$

AMEMIYA, ZELLNER et d'autres ont proposé de choisir pour l'estimateur $\widehat{DMC}$,

$$S_N^{DMC} = X (X'X)^{-1} X' \quad \text{(en général } S_N^{DMC} = H (H'H)^{-1} H')$$

où X est un N - échantillon des variables exogènes du système (ou de H un autre ensemble de variables) supposées linéairement explicatives des endogènes, à un résidu près.

$$X = \begin{pmatrix} x_{11} & \cdots & x_{s1} \\ x_{1N} & \cdots & x_{sN} \end{pmatrix}$$

AMEMIYA a démontré le théorème suivant :

Théorème 4 - (AMEMIYA)

Dans un $M_{GI}$ il existe un estimateur $\widehat{DMC}$ (minimisant $L_N$ définie ci-dessus).

Alors,

- $\hat{\theta}$ est convergent
- quand $N \to \infty$, $\sqrt{N}\,(\hat{\theta} - \theta_o)$ tend vers une distribution,

$$\mathcal{N}\left(0, \lim \left[\frac{1}{N} (G_i')_o \; S_N \; (G_i)_o\right]^{-1}\right)$$

avec,

$$G_i' = \left(\frac{\partial g_i}{\partial \theta'}\right)' = \frac{\partial g'}{\partial \sigma} = \begin{pmatrix} \frac{\partial g_{i1}}{\partial \theta_1} & \cdots & \frac{\partial g_{iN}}{\partial \theta_1} \\ \cdots & \cdots & \cdots \\ \frac{\partial g_{i1}}{\partial \theta_D} & \cdots & \frac{\partial g_{iN}}{\partial \theta_D} \end{pmatrix}$$

avec de plus, les hypothèses fixées en 2.3. dans la présentation du modèle et les conditions supplémentaires :

$k_o$ - l'espace des paramètres est compact + $\{i_o\}$

$k_1$ - $\exists \lim \left(\frac{1}{N} X'X\right)$, non singulière

$k_2$ - $\frac{1}{N} G'_i \cdot X$ converge en probabilité, ainsi que $\frac{1}{N} \frac{\partial G_i'}{\partial \theta_j} \cdot X$ ($k_3$)

toutes les deux uniformément, vers des matrices constantes, de rang D, $\forall j = 1, \ldots D$ (on rappelle $\theta \in R^D$).

Comme pour le théorème 3, la démonstration s'opère à partir d'un développement en série de TAYLOR. Elle ne fait pas intervenir la loi de U, il n'y a qu'une lointaine analogie entre les hypothèses $k_1$, $k_2$, $k_3$ et $j_1$, $j_2$, $j_3$ nécessaires pour le théorème 3. Alors, la minimisation de $L_N$ par la méthode de GAUSS-NEWTON conduit à la formule itérative pour la $k^{\text{ième}}$ itération :

$$\hat{\theta}_{(k)} = \hat{\theta}_{(k-1)} + (G_i' \cdot S_N G_i)^{-1} G_i' S_N f_i = \hat{\theta}_{(k-1)} + (\widehat{V(\theta)}^{-1} * \overrightarrow{\text{Grad}}\, L_N)_{k-1}$$

les valeurs de $f_i$ et $G_i$ étant prises en $\hat{\theta}_{(k-1)}$.

## 3.2 - propriétés asymptotiques comparées des estimateurs $\widehat{DMC}$ et $\widehat{MVIL}$

On rappelle quelques résultats sur l'estimateur du Maximum de vraisemblance à information limitée ($\widehat{MVIL}$). On étudie chaque équation du système,

$$Y_i = g_i(Y^*, X^*, \theta) + U_i \text{ notée } f_i(Y, X, \theta) = U_i$$

$$\text{où } Y_i = g(Z, \theta) + U_i$$

En général, $Y^*$ ensemble des endogènes explicatives dans cette équation est lui même supposé expliqué linéairement sous la forme du modèle,

$$Z = X.\Pi + V \qquad \text{avec } Z = (Y^*, X^*)$$

L'ensemble des variables expliquant $Z$ est généralement X mais ce peut être aussi un autre ensemble de variables H qui d'après le modèle expliquerait linéairement chaque $Y^*$ dans chaque équation.

Si les $U_i$ sont normalement distribuées, on construit la fonction de vraisemblance et l'on cherche à la Maximiser (ou son logarithme dans sa partie non constante) :

$$L = -\frac{N}{2}\,\mathrm{Log}|\Omega| - \frac{1}{2}\,\mathrm{tr}\,(\Omega^{-1}S) \text{ avec } S = \begin{pmatrix} U'U & U'V \\ V'U & V'V \end{pmatrix}$$

comme $\frac{\partial L}{\partial \Omega} = 0$, à l'optimum

$$L_o = -\frac{N}{2}\,(\log U'M_V U + \log |V'V|)$$

avec pour $M_V$ (resp $M_u$) $= 1 - V(V'V)^{-1}V'$

de plus, $\frac{\partial L}{\partial \Pi} = 0$, on a ainsi à chercher $\hat{\theta}$ qui maximise L. Par itérations, on calcule d'abord,

$$\hat{\Pi} = (X'X)^{-1} X'Z \quad \text{puis, } \hat{V} = Z - X\hat{\Pi}$$

soit $\hat{\theta}$ la valeur de $\theta$ maximisant $L_o$, alors $\hat{U} = y_i - g(\hat{\theta})$

Comme à l'optimum $\frac{\partial L}{\partial \Pi} = 0$, $\Pi = (X'M_u X)^{-1} X' M_u Z$ ; ayant $\hat{u}$ on a ainsi $\hat{\Pi}$. On itère alors la procédure jusqu'à obtenir $\theta$ convergé. On peut montrer que asymptotiquement Var ($\widehat{MVIL}$) < Var ($\widehat{DMC}$).

## 3.3. l'estimateur $\widehat{TMC}$ (triples Moindres Carrés, non linéaires).

Il s'agit de construire un estimateur qui, à la différence de l'estimateur

$\widehat{DMC}$, soit à information totale et prenne en compte $\Omega$, matrice des variances - covariances résiduelles.

Tout estimateur à distance minimale, pour l'ensemble des équations, minimise : $L_N = (y - g)' \, S_N \, (y - g)$ ; il est alors possible d'envisager pour l'estimateur $\widehat{TMC}$ de prendre

$$S_N = \hat{\Omega}^{-1} \otimes S_N^{DMC}$$

Théorème 4' -

Généralisation du théorème 4 à un système à information totale sous les hypothèses $k_o$ à $k_3$ généralisées avec en plus :

$$k_4) \quad \left[\frac{X'U_1}{\sqrt{N}}, \ldots, \frac{X'U_k}{\sqrt{N}}\right] \to \mathcal{N}(0, \Omega \otimes (\lim \frac{1}{N} X'X))$$

Dans un $M_{GI}$, il existe un estimateur $\widehat{TMC}$, (minimisant $L_N$ généralisé en 3.3)

- $\hat{\theta}$ est convergent
- quand $N \to \infty$, $\sqrt{N}\,(\hat{\theta} - \theta_o)$ tend vers une distribution,

$$\mathcal{N}(0, [\frac{1}{N} G_o' \, S_N \, G_o]^{-1}) \quad (G_o \text{ valeur de } G \text{ à la vraie valeur } \theta_o)$$

$G_i'$ a été définie en 3.1 et $G' = \begin{pmatrix} G_1' & 0 \\ 0 & G_2' \ldots \\ \vdots & \end{pmatrix}$ et $G = \begin{pmatrix} G_1 & 0 \ldots \\ 0 & G_2 \ldots \\ \vdots & \end{pmatrix}$

Nous généraliserons aussi $\widehat{DMC}$ en posant,

$$(11) \quad S_N = (\hat{\Omega} \otimes I)^{-1/2} \, S_N^* \, (\hat{\Omega} \otimes I)^{-1/2} \quad \text{avec } S_N^* = X^*(X^{*\prime} X^*)^{-1} X^{*\prime}$$

$X_N^*$ étant une matrice des variables exogènes, par exemple,

$$X_N^* = \begin{pmatrix} X & 0 & \ldots & 0 \\ 0 & X & \ldots & 0 \\ 0 & 0 & X \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & X \end{pmatrix}$$

Dans ce cas, la limite de la matrice des variances covariances de $\widehat{TMC}$ est inchangée, et le théorème 4' est valide. La minimisation de $L_N$ par la méthode de GAUSS-NEWTON nous conduit à la formule itérative suivante pour la $k^{ième}$ itération,

$$\hat{\theta}_{(k)} = \hat{\theta}_{(k-1)} - [\overline{G}' \, (\hat{\Omega}^{-1} \otimes I) \, G]^{-1} \, \overline{G}' \, (\hat{\Omega}^{-1} \otimes I) \, f \text{ où } f, G, \text{ et } \Omega \text{ sont estimés}$$

à partir de $\hat{\theta}_{k-1}$ avec,

$$\bar{G}' = \begin{pmatrix} \bar{G}'_1 & 0 \ldots \\ 0 & \bar{G}'_2 \\ \vdots & \end{pmatrix} \qquad \text{et} \qquad \bar{G}'_i = \begin{pmatrix} \frac{1}{N} & \sum_{p} \frac{\partial g_{in}}{\partial \theta_1} \\ & \vdots \end{pmatrix}$$

<u>3.4 - propriétés asymptotiques comparées de $\widehat{TMC}$ et de $\widehat{MV}$</u>

En 3.3 la matrice des variances covariances asymptotiques de $\widehat{TMC}$ a été donnée par,

$$\widehat{V(\theta)} = \left[\lim \frac{1}{N} G'_o \, S_N \, G_o\right]^{-1}$$

la limite inférieure sera l'inverse de,

$$\lim\left(\frac{1}{N} \bar{G}'_o \, \Lambda \, \bar{G}_o\right) \quad \text{avec} \quad \Lambda = \Omega^{-1} \otimes I$$

cette limite est atteinte dans notre forme (11) avec,

$$X^* = \hat{\Lambda}^{-1/2} \, \bar{G}, \quad \text{où } \theta \text{ contenu dans } G \text{ est une estimation convergente.}$$

L'estimateur ainsi obtenu sera appelé estimateur $\widehat{TMC}$ optimal ($\widehat{TMCO}$) ; il est plus difficilement calculable. Pour comparer son efficacité avec celle de $\widehat{MV}$ un lemme d'AMEMIYA est nécessaire.

<u>lemme 0</u> (AMEMIYA) - Soient $u_1 \ldots u_n$ des v.a. conjointement normales centrées et $h(u_1 \ldots u_n)$ fonction telle que $E(h)$ et $E\left(\frac{\partial h}{\partial u_i}\right)$ soient finies. Alors,

$$E(h(u_j)) = \sum_{i=1}^{n} E\left(\frac{\partial h}{\partial u_i}\right) \sigma_{ji}$$

$\sigma_{ij}$ étant la covariance entre $u_i$ et $u_j$

La démonstration de ce lemme utilise l'hypothèse $j_2$ préliminaire à la démonstration du théorème 3, qui conduit à,

$$(12) \quad \frac{1}{\sqrt{N}} \left(\frac{\partial L}{\partial \theta_i}\right)_o \underset{\text{LD}}{=} p_{i1} + p_{i2} \quad \text{où LD est la distribution limite de L}$$

$$p_{i1} = \frac{1}{\sqrt{N}} \sum \left(\frac{\partial f'_i}{\partial u_i} - f'_i \, u' \, \sigma^i\right) \quad \text{où } f'_i = \frac{\partial f_i}{\partial \theta_i}$$

et $\sigma^i$ la ième colonne de $\Omega^{-1}$

$$p_{i2} = \lim \frac{1}{N} \quad \Sigma \quad E\,(\frac{\partial f'_i}{\partial u}) \;.\; \frac{1}{\sqrt{N}}\, \Sigma\,(uu' - \Omega) \qquad \sigma^i$$

Théorème 5 - l'estimateur $\widehat{TMCO}$ est moins efficace que $\widehat{MV}$

démonstration : le lemme 1, associé à la condition $j_2$, permettra d'écrire,

$$E\left\{[\frac{\partial f'_i}{\partial u_i} - f'_i\, u_i \quad \sigma^i]\; \sigma^{j'} \quad u\; E(f'^T_j)\right\} = - \bar{s}_{ij}\; E(f'_i\; f'^T_j) \;\; (\text{noté} = - \bar{s})$$

Pour éviter toute confusion, ayant noté $f'_i$ la matrice des dérivées de $f_i$ on notera $f'^T_i$ sa transposée ; de même,

$$E\left\{E(f'_i)\; u' \quad \sigma^i [\frac{\partial f'_j}{\partial u_j} - f'_j\; u' \;\; \sigma^j]\right\} = - \bar{s}$$

$$E \quad \{(E(f'_i)\; u' \quad \sigma^i \quad \sigma^j \quad u\; E\,(f'_j)\} = \bar{s}$$

$$\text{et } E\; [(uu' - \Omega)\; \sigma^i\; \sigma^j \quad u\; E\,(f'_i)] = 0$$

$$E\; [E\,(f'_i)\; u'\; \sigma^i\; \sigma^{j'} \;\; (uu' - \Omega)] \quad = 0$$

On pose,

$$S_3 = \lim \quad E\; [\; p_{i1} + p_{i2} + \frac{\Omega}{\sqrt{N}}\, E\,(f'_i)\; u'\; \sigma^i]\;\; [p'_{j1} + p'_{j2} + \frac{\Omega}{\sqrt{N}}\, \sigma^{j'} \quad u\; E(f'_j)\;]$$

Il est évident que l'on a $S_3 \geq 0$

On pose,

$$S_1 = [\lim \quad V\; (\hat{\theta})_{\widehat{MV}}]^{-1} \qquad\qquad S_2 = [\lim\; V(\hat{\theta})_{TMCO}]^{-1}$$

Avec (12) et les espérances calculées plus haut, il vient $S_3 = S_1 - S_2$, donc $S_1 \geq S_2$

Ainsi dans les estimateurs à information totale, celui du maximum de vraisemblance est plus efficace que $\widehat{TMC}$, comme en information limitée $\widehat{MVIL}$ est plus efficace que l'estimateur $\widehat{DMC}$. Malheureusement, la convergence des estimateurs du maximum de vraisemblance est liée à la Normalité de U. De plus, les résultats sur l'efficacité sont asymptotiques.

Théorème 6 -

L'estimateur $\widehat{TMC}$ est plus efficace que $\widehat{DMC}$. En effet,

$$\mathrm{Var}\ (\widehat{TMC}) = [\frac{1}{N}\ G'_o\ S_N\ G_o]^{-1} \quad \text{quand } N \to \infty$$

$$\mathrm{Var}\ (\widehat{DMC}) = [\frac{1}{N}\ (G'_i)_o\ S_N\ (G_i)_o]^{-1}$$

La différence des deux matrices $G'_o\ S_N\ G_o - (G'_i)_o\ S_N\ (G_i)_o$ est une matrice semi-définie positive donc,

$$\mathrm{Var}\ (\widehat{MVPI}) < \mathrm{Var}\ (\widehat{TMC}) < \mathrm{Var}\ (\widehat{DMC})$$

3.5 - application économique (suite)

On pourra étudier le modèle de l'illustration originale décrite en 2.2.

$$P = \alpha_1\ 10^{\alpha_2 X_1}\ L^{\alpha_3}\ K^{1-\alpha_3}$$

ou

$$Y_1 = \alpha_1\ 10^{\alpha_2 X_1}\ Y_2^{1-\alpha_3} \quad \text{et}$$

$$Y_2 = \frac{X_2}{\alpha_3} + U_2$$

| Estimation $\widehat{DMC}$ | Modèle U. S. A. 1959-1972 | idem | Modèle France 1959-1972 | idem |
|---|---|---|---|---|
| | Convergence achevée à la 27e itération | T.M.C. | Convergence achevée à la 2e itération | T.M.C. |
| 1ère équation $\hat{\alpha}_1$ | 4,825<br>(0,894) | 2,394<br>(0,038) | 0,98<br>(0,12) | 0,98<br>(0,01) |
| $\hat{\alpha}_2$ | 0,0173<br>(0,0018) | 0,0116<br>(0,0003) | 0,017<br>(0,0068) | 0,0117<br>(0,0005) |
| $\hat{\alpha}_3$ | 0,816<br>(0,082) | 0,511<br>(0,004) | 0,47<br>(0,16) | 0,41<br>(0,08) |
| 2ème équation $\hat{\alpha}_3$ | 0,143 | | 0,331 | |

La comparaison à partir de l'estimateur $\widehat{DMC}$ laisse évidemment apparaître une différence sensible avec l'estimation par $\widehat{MCO}$, moins sensible pour l'estimation par $\widehat{MV}$ dans le cas français ; on voit par ailleurs l'importance des méthodes numériques mises en oeuvre et leur sensibilité ; ainsi pour les données américaines, les estimations par $\widehat{DMC}$ ne convergent numériquement qu'après 27 itérations, contre deux itérations dans le cas français, dans le cas de notre méthode numérique ; d'autres méthodes numériques ont été testées et se sont révélées moins rapides. D'une façon générale on remarquera que les méthodes à information totale, qu'il s'agisse d'un maximum de vraisemblance ou des triples moindres carrés, donnent des résultats très voisins et satisfaisants du point de vue des écarts types estimés des coefficients.

C'est donc à partir des estimations par $\widehat{MV}$ ou $\widehat{TMC}$ plus que par $\widehat{DMC}$ ou $\widehat{MCO}$ que pourront être entreprises ici les comparaisons France-U.S.A. sur les fonctions de production.

## 4. UN CAS PARTICULIER : L'ESTIMATEUR DES MOINDRES CARRES GENERALISES

Les conditions d'applicabilité du théorème 3, sur l'estimateur $\widehat{MVPI}$ sont très lourdes à vérifier en cas de système indépendant. L'estimateur $\widehat{MCG}$ défini en 2.3 évoqué par MALINVAUD et BARNETT pourra sous certaines conditions être asymptotiquement efficace et convergent quelque soit U.

Soit un $M_{GN}$ d'équations $y_i = g_i(X,\theta) + U$ défini comme précédemment mais avec équations indépendantes.

L'estimateur obtenu au théorème (2') est alors aussi asymptotiquement Normal et efficace ; 3 lemmes sont préliminaires à cette démonstration.

<u>lemme 4</u> - Si $X_n$ et $Y_n$ sont des suites de v.a. avec,

$$0 \leq |X_n| \leq Y_n \quad \text{et } y_n = o_p(1)$$

$$\text{alors } X_n = o_p(1)$$

<u>lemme 5</u> - (BARNETT, MALINVAUD)

$$\text{soit } \theta_N^+ = \theta_o^+ + o_p(1)$$

$$\text{alors, } N\, B_N^{-1} = I^{-1}(\theta_o^+) + o_p(1)$$

avec $B_N = - \frac{\partial^2 L}{\partial\theta^+ \partial\theta^+}$, et $a_N(\theta^+) = \frac{\partial L}{\partial\theta^+}$

$I_N(\theta^+) = - E_\theta \frac{\partial^2 L}{\partial\theta^+ \partial\theta^+}$ , matrice d'information et $I = \lim I_N$ quand $N \to \infty$

<u>Lemme 6</u> - (BARNETT, MALINVAUD)

$$\frac{1}{\sqrt{N}} a_N (\theta_o^+) \to \mathcal{N}(0, I(\theta_o^+)) \text{ quand } N \to \infty$$

<u>Théorème 7</u> - (efficacité et Normalité asymptotique)

si $\hat\theta^+$ est un estimateur $\widehat{MV}$ de $\theta_o^+$, alors

$$\sqrt{N} (\hat\theta^+ - \theta_o^+) \to \mathcal{N}(0, I^{-1}(\theta_o^+)) \quad \text{quand} \quad N \to \infty$$

*Par le théorème de la moyenne multivariée*

$$W_i = (\frac{1}{N} \frac{\partial L}{\partial\theta_i})_{\hat\theta^+} = \frac{1}{N} (\frac{\partial L}{\partial\theta_i})_{\theta_o^+} + \frac{1}{N} (\frac{\partial^2 L}{\partial\theta' \partial\theta_i})_{\theta^*} (\hat\theta^+ - \theta_o^+)$$

*avec* $\theta^*$ *entre* $\hat\theta^+$ *et* $\theta_o^+$

où,

$$W_i = \frac{1}{N} a_{N_i} (\theta_o^+) - \frac{1}{N} b'_{N_i} (\theta^*) (\hat\theta^+ - \theta_o^+)$$

$b'_{N_i}$ étant la $i^{\text{ème}}$ ligne de $B_N$

si $\hat\theta$ est un estimateur $\widehat{MV}$, $W = 0$, alors,

$$(\hat\theta^+ - \theta_o^+) = N B_N^{-1}(\theta^*) \frac{1}{N} a_N (\theta_o^+)$$

or, $0 \leq |\theta^* - \theta_o^+| \leq |\hat\theta - \theta_o^+|$ avec $|\hat\theta - \theta_o| = o_p$ (1)

d'après les lemmes [illegible] et 5, [illegible]ent,

$$\theta^* = \theta_o^+ + o_p (1)$$

$$N B_N^{-1}(\theta^*) = I^{-1}(\theta_o^+) + o_p (1)$$

d'où $\sqrt{N} (\hat\theta^+ - \theta_o^+) = [I^{-1}(\theta_o^+) + o_p(1)] \frac{1}{\sqrt{N}} a_N (\theta_o^+)$

$$= I^{-1}(\theta_o^+)\left(\frac{1}{\sqrt{N}}\, a_N(\theta_o^+)\right) + o_p(1) \quad \text{(lemme 6)}$$

Par le lemme 6 et le théorème de Slutsky, on a alors démontré ce théorème. On a en corrolaire,

$$\sqrt{N}\,(\hat{\theta} - \theta_o) \to_{\mathcal{L}} \mathcal{N}(0,\ I^{-1}(\theta_o^+)) \text{ quand } N \to \infty$$

On peut montrer que les éléments de la diagonale de $B_N^{-1}(\hat{\theta}^+)$ représentent asymptotiquement $\widehat{Var}\,(\theta^+)$.

Par ailleurs, dans les $M_{GI}$, une recherche similaire reste à faire pour la classe dite des M-estimateurs.

# BIBLIOGRAPHIE


[0 ] BARNETT W. - "Maximum Likelihood and Iterated Aitken Estimation of Nonlinear Systems of Equations"- JASA juin 1976 - 354.360 -

[0'] AMEMIYA T. - "The Maximum Likehood and the Nonlinear Three-Stage Least-Squares Estimator in the General Nonlinear Simultaneous Equation Model"- Econométrica 1977 - 955.968 -

[1 ] AMEMIYA T. - "L'estimation des modèles à équations simultanées non linéaires"- Cahiers du Séminaire d'Econométrie - 1978-1.

[2 ] BERNDT E.K., H.B. HALL, R.E. HALL and J.A. HAUSMAN - "Estimation and inference in Nonlinear structural models" - 1974 - Annals of Economic and Social Measurement 3, 653.666 -

[3 ] EISENPRESS H. and J. GREENSTADT - "The Estimation of Nonlinear Econometric systems"- 1966 - Econometrica 34 - 851.861 -

[4 ] GALLANT A.R. - "Seemingly unrelated nonlinear regressions" 1975 a, Journal of Econometrics 3 - 35.50 -

[5 ] GALLANT A.R. et HOLLY A. - "Statistical Inference in an implicit non linear simultaneous equation model" - ronéo 78.

[6 ] JENNRICH R.L. - "Asymptotic properties of non-linear least squares estimators" 1969, Annals of Mathematical Statistics 40, 633.643 -

[7 ] JORGENSON D.W, and J. LAFFONT - "Efficient estimation of nonlinear simultaneous equations with additive disturbances", 1974, Annals of Economic and Social Measurement 3, 615.640 -

[8 ] MALINVAUD E. - "The Consistency of nonlinear regressions" 1970, Annals of Mathematical Statistics 41, 956.969 -

[9 ] MALINVAUD E. - "Méthodes statistiques de l'économétrie" 1978, Dunod, Paris.