REVUE DE STATISTIQUE APPLIQUÉE

Y. ARAGON

Sur l'estimation des probabilités de transition d'une chaîne de Markov à partir des données marginales

Revue de statistique appliquée, tome 20, nº 2 (1972), p. 79-94
<http://www.numdam.org/item?id=RSA_1972__20_2_79_0>

© Société française de statistique, 1972, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

SUR L'ESTIMATION DES PROBABILITÉS DE TRANSITION D'UNE CHAINE DE MARKOV A PARTIR DES DONNÉES MARGINALES

Y. ARAGON
Laboratoire de Statistique (Université de Toulouse)

I - INTRODUCTION

L'estimation des probabilités de transition dans une chaîne de Markov, à partir d'observations portant sur les transitions (micro-observations) a été traitée par Anderson et Goodman [1]. Cependant l'estimation de ces probabilités à partir des données marginales (macro-observations) n'a pas reçu de solution satisfaisante.

C'est ce problème qui est étudié ici. Le but de cet article est essentiellement pratique. Aussi n'y figurent que peu d'éléments de justification. Pour une étude plus complète on pourra se reporter à [2]. Par contre, la solution numérique du problème a été détaillée pour permettre à tout praticien d'utiliser facilement les techniques proposées.

II - POSITION DU PROBLEME

Pour un utilisateur, une chaîne de Markov est un modèle dynamique discret permettant d'analyser des séries chronologiques (voir Telser [12] pour un exposé des applications économiques classiques). Ce modèle repose sur l'hypothèse suivante :

"La probabilité pour un individu d'être dans un état j donné à un instant t, est entièrement déterminée par sa position à l'instant précédent ou d'une manière générale aux n instants consécutifs précédents ; n est l'ordre de la chaîne".

Les paramètres d'une chaîne de Markov sont les probabilités de transition p_{ij}, formant la matrice P.

p_{ij} est la probabilité pour un individu quelconque de se trouver dans l'état j à un instant sachant qu'il occupait l'état i à l'instant précédent. Il y a r états $(i,j = 1,\ldots,r)$. Les probabilités marginales $q_j(t)$ qu'un individu quelconque soit dans l'état j à l'instant t vérifient la relation de récurrence :

$$q_j(t) = \sum_{i=1}^{r} p_{ij}\ q_i(t-1) \quad . \quad \begin{array}{l} t=1,\dots,T \\ j=1,\dots,r \end{array}$$

On dispose, pour estimer les p_{ij}, d'échantillons extraits de cette population. Dans la situation classique, on observe les $n_{ij}(t)$, nombre d'individus passés de l'état i à l'instant t-1 à l'état j en t et pour des temps d'observations consécutifs en nombre T+1, et notés de 0 à T.

La présente étude est centrée sur le cas où les $n_{ij}(t)$ sont inconnus et où seules sont disponibles les marges $n_j(t)$, c'est-à-dire les nombres d'individus présents dans un état à un instant. On a évidemment :

$$n_i(t-1) = \sum_{j=1}^{r} n_{ij}(t) \qquad t=1,\dots,T \qquad i=1,\dots,r \tag{1}$$

$$n_j(t) = \sum_{i=1}^{r} n_{ij}(t) \qquad t=0,\dots,T \qquad j=1,\dots,r.$$

La taille de l'échantillon observé au temps t est $N(t) = \sum_{j=1}^{r} n_j(t)$; dans les cas où cette taille est indépendante de t on la note N. Remarque : on parlera indifféremment de marges, de données marginales, de macro-observations, de données agrégées.

III - DIFFICULTES DANS L'APPLICATION DE LA METHODE DU MAXIMUM DE VRAISEMBLANCE, TELLE QU'ELLE EST UTILISEE EN [1] :

Dans la situation classique, le même échantillon est observé à chaque instant, et on connaît les transitions. Quand la répartition initiale des individus est certaine la densité conjointe des observations, se note :

$$L(n(1),\dots,n(T)/n(0)) = \prod_{t=1}^{T} \left\{ \prod_{i=1}^{r} \left[\frac{n_i(t-1)!}{\prod_{j=1}^{r} n_{ij}(t)!} \prod_{j=1}^{r} p_{ij}^{n_{ij}(t)} \right] \right\}.$$

Quand les micro-observations $n_{ij}(t)$ sont connues, on obtient à partir de cette expression, les estimateurs M. L. des p_{ij}.

Si les $n_{ij}(t)$ sont inconnus et que seuls les $n_j(t)$, données marginales sont disponibles, on peut formuler la densité conjointe des observations $n_j(t)$.

a) Probabilité conjointe de $n(t) = (n_1(t),\dots,n_r(t))$ conditionnée par la connaissance de $n(t-1)$.

Il faut dénombrer toutes les façons possibles de passer de l'état i en t-1, à l'état j en t, et les affecter de leur probabilité de réalisation. Or il y a autant de façons de passer de i en t-1 à j en t, qu'il y a de $n_{ij}(t)$, vérifiant les 3 conditions :

$$\begin{cases} 0 \leqslant n_{ij}(t) \leqslant n_i(t-1) & i,j=1,\dots,r \\ \sum_{i=1}^{r} n_{ij}(t) = n_j(t) & j=1,\dots,r \\ \sum_{j=1}^{r} n_{ij}(t) = n_i(t-1) & i=1,\dots,r\,. \end{cases} \tag{3}$$

D'autre part $n_{ij}(t)$ est distribuée multinomialement de paramètres : $(n_i(t-1);p_{i1},\dots,p_{ir})$. La probabilité d'observer le vecteur $n(t)$ conditionnée par la connaissance de $n(t-1)$ s'écrit donc :

$$f(n_1(t),\dots,n_r(t)|n(t-1) = \sum_{n_{ij}(t)} \prod_{i=1}^{r} \left\{ n_i(t-1)!\ \prod_{j=1}^{r} \frac{p_{ij}^{n_{ij}(t)}}{n_{ij}(t)!} \right\}$$

avec $n_{ij}(t)$ vérifiant les conditions (3).

C'est le produit de convolution de r densités multinomiales.

b) On peut maintenant écrire la densité conjointe des observations, la répartition initiale $n(0)$ étant supposée connue.

$$L(n(1),\dots,n(T)|n(0)) = \prod_{t=1}^{T} \left\{ \sum_{n_{ij}(t)} \prod_{i=1}^{r} \left[n_i(t-1)! \prod_{j=1}^{r} \frac{p_{ij}^{n_{ij}(t)}}{n_{ij}(t)!} \right] \right\} \tag{4}$$

les $n_{ij}(t)$ vérifiant (3).

Cette expression est impraticable pour rechercher des estimateurs M. L. des p_{ij}. C'est pourquoi d'autres méthodes d'estimations convenables sous certaines hypothèses ont été proposées.

IV - AUTRE APPROCHE DU PROBLEME

Plusieurs types d'hypothèses sur le modèle sont possibles parmi lesquelles les deux suivantes s'introduisent simplement. Pour chacune d'elles nous donnons les estimateurs obtenus sans tenir compte des contraintes d'inégalités.

Ces estimateurs n'appartiennent pas nécessairement à l'espace paramétrique Ω_*

$$\Omega_* = \{p_* = (p_{11},p_{21},\dots,p_{r1},p_{12},\dots,p_{rr-1}\}' \ (p_{ij} \geqslant 0(j=1,\dots)r-1),$$
$$1 - \sum_{j=1}^{r-1} p_{ij} \geqslant 0),\ i=1,\dots,\}r.$$

C'est pourquoi on indiquera à la section V, une procédure simple et rapide permettant d'obtenir des estimateurs appartenant à Ω_*.

A - Situation où les marges sont directement obtenues (en abrégé - situation M).

1 - Etant données que seules les marges sont utilisées, on peut envisager des échantillons où elles sont directement obtenus, sans l'intermédiaire des transitions. Ce sera notamment le cas, si on prélève à tous les instants d'observation consécutifs, dans la population markovienne d'ordre 1, un nouvel échantillon tiré indépendamment du précédent au moment du sondage. En effet chacun de ces échantillons est composé d'individus présents dans l'état i à l'instant t. Il est distribué suivant une loi multinomiale (N(t) ; $q_1(t),\ldots,q_r(t)$. On a bien obtenu ainsi des marges $n_i(t)$ $(i=1,\ldots,r)$, distribuées multinomialement. La densité conjointe de ces observations s'écrit :

$$f(n(1),\ldots,n(T)) = \prod_{t=1}^{T} N(t)! \prod_{i=1}^{r} \frac{q_i(t)^{n_i(t)}}{n_i(t)!} \tag{5}$$

Les paramètres à estimer p_{ij}, apparaissent dans cette expression par l'intermédiaire des relations

$$q_j(t) = \sum_{i=1}^{r} p_{ij}\, q_i(t-1) \qquad t=1,\ldots,T \qquad j=1,\ldots,r \tag{6}$$

ou

$$q_j(t) = \sum_{i=1}^{r} p_{ij}^{(t)}\, q_i(0) \tag{7}$$

$p_{ij}^{(t)}$ est l'élément ij de P^t. Si $q_i(0)$ est inconnue on peut l'estimer par $\frac{n_i(0)}{N(0)}$, et l'expression (5) est remplacée par

$$g(n(1),\ldots,n(T)\,|\,n(0)) = \prod_{t=1}^{T} N(t)! \prod_{i=1}^{r} \frac{q_i(t)^{n_i(t)}}{n_i(t)!} \; . \tag{8}$$

La recherche des estimateurs M. L. des p_{ij} à l'aide des relations (5) et (7) ou (8) et (7) est encore trop compliquée. Lee et Collaborateurs [6] ont donc utilisé au lieu des relations exactes (8), des estimations de ces expressions, obtenues en remplaçant $q_j(t) = \sum_{i=1}^{r} p_{ij}\, q_i(t-1)$ par

$$\rho_j(t) = \sum_{i=1}^{r} p_{ij} \frac{n_i(t-1)}{N(t-1)} \tag{9}$$

Quand $N(t) \to +\infty$ pour tout t, $\rho_j(t)$ converge en probabilité vers $q_j(t)$.

Au lieu de (8) on a :

$$h(n(1),\ldots,n(T)\,|\,n(0)) = \prod_{t=1}^{T} N(t)! \prod_{i=1}^{r} \frac{\rho_i(t)^{n_i(t)}}{n_i(t)!} \tag{10}$$

où $\rho_i(t)$ est défini par (9).

Cette expression n'est pas la vraisemblance de l'échantillon dont on vient de décrire l'obtention.

(cf. [2] pour une interprétation de (10) comme vraisemblance).

2 - Estimateurs sans contraintes dans cette situation

On montre [6] que les estimateurs obtenus en maximisant l'expression (10) sont solutions de l'équation :

$$X_*' \Sigma_*^{-1} W_* - X_*' \Sigma_*^{-1} X_* p_* = 0 \tag{11}$$

$$\left.\begin{array}{l} \text{où} \quad \Sigma_* = \begin{bmatrix} \Sigma_{11} \cdots\cdots\cdots\cdots \Sigma_{1,r-1} \\ \Sigma_{r-1,1} \cdots\cdots\cdots\cdots \Sigma_{r-1,r-1} \end{bmatrix} \\ \text{et } \Sigma_{ij} = \begin{bmatrix} \dfrac{\rho_i(1)\ (\delta_{ij} - \rho_j(1))}{N(1)} \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \dfrac{\rho_i(T)\ (\delta_{ij} - \rho_j(T))}{N(T)} \end{bmatrix} \end{array}\right\} \tag{12}$$

$$\left.\begin{array}{ll} W_* = [W_1, \ldots, W_{r-1}]' & p_* = [p_1', \ldots, p_{r-1}']' \\ W_i = [W_i(1), \ldots, W_i(T)]' & p_i = [p_{1i}, \ldots, p_{ri}]' \\ X_i = X = \begin{bmatrix} W_1(0) \cdots\cdots W_r(0) \\ \cdots\cdots\cdots\cdots \\ W_1(T) \cdots\cdots W_r(T) \end{bmatrix} & \end{array}\right\} \tag{13}$$

Σ_* étant inconnue, on la remplace par un estimateur $\hat{\Sigma}_*$.

p_* solution de (11) est aussi l'estimateur G. L. S. qu'on aurait obtenu en minimisant :

$$(W_* - X_* p_*)' \Sigma_*^{-1} (W_* - X_* p_*)$$

B - Situation où les marges sont des agrégations de transitions (en abrégé situation T).

1 - Anderson et Goodman envisagent les transitions d'un seul échantillon d'individus qui transitent tous de l'observation initiale à l'observation finale, par les états de la chaîne de Markov. Pour trouver les estimateurs des p_{ij}, ils doivent connaître les transitions entre deux instants consécutifs (au moins). L'échantillon suivant le processus évolue exactement comme le suggère la définition d'une chaîne de Markov et c'est ainsi qu'habituellement on engendre artificiellement un échantillon de cette population.

2 - Estimateurs par les moindres-carrés sans contraintes d'inégalités dans cette situation (situation T) :

On peut essayer d'estimer les p_{ij} par une méthode de moindres carrés (L.S.) sur un système d'équations linéaires. On peut remplacer les relations (6) par les relations

$$\frac{n_j(t)}{N} = \sum_{i=1}^{r} p_{ij} \frac{n_i(t-1)}{N} + u_j(t)$$

ou si $W_j(t) = \frac{n_j(t)}{N}$

$$W_j(t) = \sum_{i=1}^{r} p_{ij} W_i(t-1) + u_j(t) \qquad j=1,\ldots,r. \quad t=1,\ldots,T , \tag{14}$$

$u_j(t)$ est un aléa centré dont la distribution reste à préciser. Le modèle de régression introduit par les relations (14) peut être considéré comme une régression conditionnelle de $W_j(t)$ sur $W_i(t-1)$. On posera alors :

$$E(u_i(s)|W(s) \quad \text{pour } s < t) = 0$$

C'est une convention cohérente avec le mode d'obtention des marges, à partir du même échantillon à tous les instants. Les moments centrés d'ordre 2 d'un produit de convolution étant la somme des moments correspondants des lois convoluées, la matrice des covariances du produit de convolution (4) est donc :

$$= \left\| \sum_{j=1}^{r} n_j(t-1) \ (\delta_{ih} - p_{jh}) \ p_{ji} \right\|_{ih} \tag{15}$$

On cherche en fait les moments d'ordre 2 des $u_j(t)$ définis par (12). Sous l'hypothèse précédemment introduite qu'à chaque instant, les $W_i(s)$ pour $s < t$ sont certains, on passe de $W_i(t)$ à $u_i(t)$ par une translation certaine, on montre que $\mathrm{cov}(u_j(s), \ u_k(t)|W_i(s), i=1,\ldots,r, \ s=0,\ldots,t-1) = 0$.

De (15) et de $W_j(t-1) = \frac{n_j(t-1)}{N}$, on déduit [2]

$$\mathrm{cov}(u_h(t), u_i(t)|W_k(s), k=1,\ldots,r, \ s=0,\ldots,t-1) = \sum_{j=1}^{r} \frac{n_j(t-1)}{N^2} (\delta_{ih} - p_{jh}) \ p_{ji}$$

Pour un échantillon de taille assez grande, on peut donc approcher la distribution des $u_i(t)$ $(i=1,\ldots,r)$ par une distribution multinormale centrée, de matrice des covariances

$$\Lambda_t = \left\| \sum_{j=1}^{r} \frac{n_j(t-1)}{N^2} (\delta_{ih} - p_{jh}) \ p_{ji} \right\|$$

On peut exprimer les équations (14) comme un système d'équations linéaires (11).

$$\begin{bmatrix} W_1 \\ \cdot \\ \cdot \\ \cdot \\ W_{r-1} \end{bmatrix} = \begin{bmatrix} X_1 \cdots\cdots \\ \cdots\cdots\cdots \\ \cdots\cdots\cdots \\ \cdots\cdots\cdots \\ \cdots X_{r-1} \end{bmatrix} \begin{bmatrix} p_1 \\ \cdot \\ \cdot \\ \cdot \\ p_{r-1} \end{bmatrix} + \begin{bmatrix} u_1 \\ \cdot \\ \cdot \\ \cdot \\ u_{r-1} \end{bmatrix} \tag{16}$$

$$W_* = X_* \, p_* + u_* \tag{17}$$

$$u_i = [u_i(1), \ldots, u_i(T)]'$$

Dans ce système, les covariances des éléments de u_* sont les covariances conditionnelles précédemment introduites. La matrice Λ_* de ces covariances n'est donc pas la matrice des covariances des éléments de u_* au sens habituel du terme. Mais comme on a convenu d'effectuer des régressions conditionnelles de W(t) sur W(t-1), on peut montrer que :

$$(W_* - X_* \, p_*)' \, \Lambda_*^{-1} \, (W_* - X_* \, p_*) = \sum_{t=1}^{T} (W(t) - P_* \, W_*(t-1))' \ldots$$

$$\ldots \Lambda_t^{-1} \, (W(t) - P_* \, W_*(t-1)) \tag{18}$$

P_* étant déduite de P en éliminant la dernière colonne $W_*(t-1)$ étant déduit de W(t-1) en éliminant son dernier élément.

Cette égalité incite à choisir la régression linéaire :

$$W_* = X_* \, p_* + u_*$$

où u_* est une v.a. centrée pour laquelle on prend comme matrice des covariances Λ_*.

$$\left.\begin{aligned} \Lambda_* &= \begin{bmatrix} \Lambda_{11} \cdots \Lambda_{1r-1} \\ \cdots\cdots\cdots\cdots \\ \Lambda_{r-1} \cdots \Lambda_{r-1,\,r-1} \end{bmatrix} \\ \Lambda_{ij} &= \begin{bmatrix} \sum_{j=1}^{r} \frac{n_j(0)}{N^2} (\delta_{ij} - p_{jh}) \, p_{ji} \cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots \sum_{j=1}^{r} \frac{n_j(T-1)}{N^2} (\delta_{ij} - p_{jh}) \, p_{ji} \end{bmatrix} \end{aligned}\right\} \tag{19}$$

La minimisation de $(W_* - X_* p_*)' \, \Lambda_*^{-1} (W_* - X_* p_*)$, donnerait les estimateurs G. L. S. (moindres-carrés généralisés), si Λ_* était connue, mais Λ_* dépend des paramètres à estimer. C'est pourquoi Λ_* est remplacée par une estimation $\hat{\Lambda}_*$ de cette matrice. Le choix de $\hat{\Lambda}_*$ est envisagé à la section V.

C - <u>Finalement les estimateurs obtenus dans les deux situations ci-dessus, sont solutions du problème</u>

$$\min \, (W_* - X_* \, p_*)' \; B^{-1} \, (W_* - X_* \, p_*) \tag{I}$$

où $B = \Lambda_*$ (situation T)

Σ_* (situation M).

Mis à part les estimateurs bayesiens évoqués en annexe, presque tous les estimateurs proposés jusqu'à présent sont des estimateurs par les moindres-carrés (généralisés), solutions du problème (I); seule la matrice B diffère d'un auteur à l'autre.

Mais on constate que les estimateurs obtenus n'appartiennent pas à l'espace paramétrique. Tous les problèmes à résoudre doivent être posés comme des problèmes de maximum sous contraintes.

Au lieu du problème (I) on doit donc résoudre :

$$\min_{p_* \in \Omega_*} (W_* - X_* p_*)' \; B^{-1} \; (W_* - X_* p_*) \Big\} \qquad \text{(II)}$$

La solution numérique du problème se trouve compliquée car on est amené à utiliser une méthode de programmation mathématique.

Dans le tableau ci-dessous on a regroupé :

- les choix faits jusqu'à présent par plusieurs auteurs
- le modèle linéaire
- la matrice de pondération : B
- la méthode de programmation sans contraintes
- la méthode de programmation mathématique.

Les comparaisons numériques permettent de croire que les méthodes les plus élaborées fournissent les meilleurs résultats.

Equation linéaire	Matrices de pondération B =	Type d'échantillons idéal pour cette situation	Méthode	Programme mathématique ($p_* \in \Omega_*$)	Référence
$W_* = X_* p_* + u_*$	diag $E(\Lambda_*)$ (1)	T	G. L. S.		9
	I		L. S.	Méthode empirique	3
	I		M. A. D. (2)	Simplex	6
	$E(\Lambda_*)$ (1)		G. L. S.		8
	Σ_*	M	G. L. S.	Méthode de Wolfe	7
	Λ_*	T	G. L. S.	Méthode de pénalisation	2
	$\sigma^2 I$		L. S.	Méthode de Wolfe	5

(1) diag $E(\Lambda_*)$ = matrice diagonale formée avec la diagonale de $E(\Lambda_*)$
$E(\Lambda_*)$ = matrice déduite de Λ_* en remplaçant

$$\sum_{j=1}^{r} \frac{n_j(t)}{N^2} (\delta_{ij} - p_{jh}) \, p_{ji} \text{ par } \sum_{j=1}^{r} \frac{E(n_j(t))}{N^2} (\delta_{ij} - p_{jh}) \, p_{ji}$$

(2) M. A. D. = Déviation absolue minimum. On minimise $|W_* - X_* p_*|' E$

où $|W_* - X_* p_*|$ est le vecteur des valeurs absolues des éléments de $W_* - X_* p_*$

E est un vecteur colonne de 1.

V - SOLUTION NUMERIQUE DU PROBLEME

Nous donnons maintenant la solution numérique que nous avons adopté. Elle est simple à programmer et peut dans certains cas recevoir une interprétation statistique.

Nous avons vu que le problème se pose ainsi

$$\min_{p_* \in \Omega_*} (W_* - X_* p_*)' \; B^{-1} (W_* - X_* p_*) \qquad \text{(II)}$$

où B est une des matrices du tableau précédent.

A - 1 - Choix de $\hat{\Sigma}_*$ (situation M). Si on ne connait pas d'estimateurs $\hat{p}_*$ de p_*, on peut prendre pour $\hat{\Sigma}_*$, la matrice obtenue en remplaçant dans Σ_* : $\rho_i(t)$ par $W_i(t)$, estimateur convergent de $\rho_i(t)$.

On peut donc calculer un $p_*^{(1)}$ en résolvant le problème B et calculer un nouveau Σ_* en remplaçant $\rho_i(t)$ par $\sum_{j=1}^{r} p_{ji} W_j(t-1)$.

Avec le nouvel estimateur de Σ_*^{-1} obtenu ainsi on peut résoudre à nouveau le problème (B) et ainsi de suite.

2 - Choix de Σ_* (situation T). Si on ne connaît pas d'estimateur $\hat{p}_*$ de p_* on peut prendre la même matrice qu'à l'étape initiale de la situation M, avec cette matrice on résout le problème (II), on obtient un p_*, on peut alors le reporter dans l'expression (19) pour obtenir un nouvel estimateur de Λ_*, et recommencer l'opération.

B - Simplifiant les notations, on doit résoudre à chaque itération le programme mathématique :

$$\left.\begin{array}{l} \min f(x) = \dfrac{1}{2} x' Qx - x'D \\ \text{pour } x = (x_1, \ldots, x_{r(r-1)})' \in \Omega_* \end{array}\right\} \qquad \text{(III)}$$

Dans cette expression :

$$Q = \begin{cases} X_* \hat{\Lambda}_*^{-1} X_* & \text{situation T} \\ X'_* \hat{\Sigma}_*^{-1} X_* & \text{situation M} \end{cases}$$

$$D = \begin{cases} 2 X'_* \hat{\Lambda}_*^{-1} W_* & \text{situation T} \\ 2 X'_* \hat{\Sigma}_*^{-1} W_* & \text{situation M} \end{cases}$$

$$p_{jk} = x_{r(k-1)+j} \begin{cases} k=1, \ldots, r-1 \\ j=1, \ldots, r \, . \end{cases}$$

La méthode la plus simple dans notre cas, semble être la méthode de pénalisation. L'idée des méthodes de pénalisation est de ramener un tel programme à la recherche d'une suite de minimums libres appartenant à Ω_*. Pour cela f(x) est corrigée par une pénalisation. Suivant le choix de cette pénalisation la méthode numérique peut recevoir une interprétation statistique.

1 - Choix d'une fonction de pénalisation

Suivant Fiacco et Mc Cormick [4] nous avons adopté

$$P(x;\varepsilon) = f(x) + \varepsilon\, L(x)$$

$$L(x) = \sum_{i=1}^{r(r-1)} \frac{1}{x_i} + \sum_{j=1}^{r} \frac{1}{1 - \sum_{k=1}^{r} x_{r(k-1)+j}} \qquad (*)$$

(Précisons que L n'est pas ici une vraisemblance).

ε est un "paramètre de perturbation" strictement positif.

ε L(x) est une pénalisation, telle que le minimum de $P(x;\varepsilon)$ appartient à Ω_*. Dans la méthode, ayant obtenu pour $x = x_0$, le minimum de $P(x;\varepsilon_0)$ où ε_0 est la valeur initiale de ε, on choisit ε_1 tel que $0 < \varepsilon_1 < \varepsilon_0$ et partant de x_0 on cherche le minimum de $P(x;\varepsilon_1)$.

D'une manière générale, on obtient x_n tel que $P(x_n;\varepsilon_n) = \min P(x;\varepsilon_n)$.

Quand $\varepsilon_n \to 0$, $\min P(x;\varepsilon_n) \to \min_{x\in\Omega_*} f(x)$

Remarque : Contrairement aux méthodes dérivées du simplexe, les méthodes de pénalisation exigent la prise en compte explicite de toutes les contraintes ($x_i \geqslant 0$).

Choix de ε_0 :

Ayant choisi un point initial x_* intérieur à Ω_*, on convient de choisir [4] :

$$\varepsilon_0 = \left.\frac{|\nabla f . \nabla L|}{|\nabla L|^2}\right|_{x=x_*}$$

Le choix de x_* ne pose pas de difficultés étant donnée la forme des contraintes.

2 - Résolution de $\min P(x;\varepsilon_n)$

Nous avons utilisé l'algorithme de Marquardt [10], mais tout algorithme de minimisation de fonction non linéaire est en principe acceptable. La convergence est assez rapide dans notre choix.

(*) On peut aussi adopter pour ε L(x) ; l'expression :

$$\sum_{i=1}^{r(r-1)} \varepsilon_i \operatorname{Log} x_i - \sum_{j=1}^{r} \varepsilon_j' \operatorname{Log} \left(1 - \sum_{k=1}^{r} x_{r(k-1)+j}\right)$$

où ε_i, ε_j' sont strictement positifs.

Ces considérations générales exposées, nous donnons, un organigramme assez détaillé, du calcul des estimateurs des probabilités de transition.

Organigramme du calcul des estimateurs

1/ Lire les observations de la chaîne.

2/ Calculer un estimateur de Σ_*^{-1} (ou Λ_*^{-1} suivant la situation).

3/ Former les termes quadratique Q et linéaire B de l'expression à minimiser à l'aide de Σ_*^{-1} (resp. Λ_*^{-1}) et des observations.

4/ Effectuer la Procédure de Minimisation Séquentielle Libre (PMSL). Soit $p_*^{(n)}$ la valeur obtenue.

5/ Calculer Σ_* (resp. Λ_*) en fonction de $p_*^{(n)}$

6/ Recommencer en 3/. Arrêter le calcul quand $|p_*^{(n)} - p_*^{(n+1)}| < \varepsilon_p$ où ε_p est strictement positif et arbitrairement petit.

Procédure de Minimisation Séquentielle Libre

1/ Définir le domaine par des inégalités : $g_i \geqslant 0$, $i \varepsilon I$

$$\Omega = \{x \; ; \; g_i \geqslant 0, \; i \in I\}$$

2/ Former une fonction pénalisée P. Ici :

$$P(x;\varepsilon_n) = \frac{1}{2} x' \, Qx - x' - B + \varepsilon_n \sum_{i \in I} \frac{1}{g_n}$$

3/ Choisir un point intérieur à Ω. Calculer la valeur $\varepsilon = \varepsilon_o$ correspondante.

4/ A la n-ième itération ($n = 0,1,\ldots$).

Calculer le minimum de P par une méthode itérative (par exemple algorithme de Marquardt), en prenant comme valeur initiale du vecteur x, le dernier point intérieur obtenu. Soit $x(\varepsilon_n)$ la valeur de x pour laquelle $P(x;\varepsilon_n)$ est minimum.

5/ Diminuer ε. Par exemple $\varepsilon_1 = \varepsilon/3$

6/ Avec $x(\varepsilon_n)$ comme point initial et $\varepsilon_{n+1} = \varepsilon_1$ recommencer en 4/. Arrêter le calcul quand $|x(\varepsilon_n) - x(\varepsilon_{n+1})| < \delta$ où δ est strictement positif et arbitrairement petit.

4 - Remarques et compléments sur la P.M.S.L.

Numéro 2 : la pénalisation choisie est du type $1/g_i$. On aurait pu choisir une pénalisation du type Log g_i.

Numéro 3 : le choix d'un point intérieur ne soulève pas de difficultés étant donné le problème. cf Fiacco et Mc Cormick [4] pour une utilisation de la P.M.S.L. dans le choix d'un point intérieur initial.

Numéro 4 : le choix de l'algorithme de Marquardt est justifié car il est rapidement convergent, mais en contre partie, eu égard à la pénalisation en $1/g_i$ adoptée, on introduit dans le calcul du Hessien des termes en ε/g_{iz} très sensibles aux erreurs d'arrondis (le vecteur x calculé sort du domaine). On peut remédier à cette situation de plusieurs façons.

$\underline{\alpha}$ En recommençant la minimisation libre à partir du dernier $x(\varepsilon)$ appartenant à Ω, obtenu, mais avec un ε_1 supérieur à celui qui donne de mauvais résultats. Ce faisant on se rapproche plus lentement de la frontière (et de la solution).

$\underline{\beta}$ En utilisant un algorithme de minimisation introduisant moins d'erreurs d'arrondis (par exemple, méthode de Fletcher-Powell) mais plus lentement convergent. Cet échange peut être fait si l'opération α recommencée plusieurs fois ne donne toujours pas de meilleurs résultats.

$\underline{\gamma}$ En donnant aux composantes de x, violant les contraintes, des valeurs proches des valeurs obtenues, mais ramenant x dans Ω. Ceci quand les moyens énoncés en α et β ne donnent pas de résultats.

$\underline{\delta}$ En choisissant une pénalisation du type Log g_i.

Numéro 5 : il est difficile de fixer a priori le choix du taux de diminution de ε. Ceci est affaire d'expérience. Nous avons adopté $\varepsilon_{n+1} = \varepsilon_n/3$ ou $\varepsilon_{n+1} = \varepsilon_n/2$.

VI - CONCLUSION

Nous donnons pour terminer des résultats d'expériences de simulations portant sur 50 échantillons, de 1000 individus transitant entre 3 états au cours de 25 périodes d'observations, suivant la situation T. Ces expériences portaient sur deux matrices de transitions et avaient pour but de comparer les estimateurs obtenus par la méthode de Lee, Judge et Zellner (M. L. ou G. L. S.) et les estimateurs introduits dans [2]. Les estimateurs introduits dans [2] apparaissent sensiblement meilleurs, pour estimer les éléments de P_2.

MATRICES

P_1			P_2		
0.35	0.35	0.3	0.9	0.05	0.05
0.4	0.3	0.3	0.0	0.9	0.1
0.3	0.35	0.35	0.01	0.01	0.98

RESULTATS

Méthode introduite dans cette étude				Méthode de Lee et collaborateurs		
Moyenne des résultats obtenus						
0.324	0.382	0.294		0.324	0.380	0.294
0.407	0.251	0.342	P_1	0.409	0.251	0.340
0.312	0.362	0.326		0.318	0.364	0.318
0.828	0.132	0.040		0.820	0.140	0.040
0.062	0.822	0.116	P_2	0.067	0.816	0.117
0.006	0.015	0.979		0.006	0.016	0.978

Carrés moyens des erreurs

0.018	0.016	0.013		0.018	0.016	0.013
0.019	0.019	0.023	P_1	0.019	0.019	0.023
0.018	0.013	0.019		0.019	0.013	0.017
0.008	0.001	0.002		0.009	0.013	0.002
0.005	0.010	0.002	P_2	0.006	0.012	0.002
0.000	0.000	0.000		0.000	0.000	0.000

ANNEXE

I - En admettant que (10) est une vraisemblance, on peut envisager des estimateurs bayesiens de p_*

Il est courant de choisir comme loi de probabilité a priori, pour un vecteur de paramètres positifs dont la somme est 1 , une loi Beta multidimensionnelle.

On notera la loi a priori du vecteur $p_i = (p_{i1}, \ldots, p_{ir})$:

$$f_i(p_i) = \frac{\Gamma(a_{i1}+\ldots+a_{ir})}{\Gamma(a_{i1})\ldots(a_{ir})} p_{i1}^{a_{i1}-1} \ldots p_{ir}^{a_{ir}-1}$$

Les a_{ij} quantifient la connaissance a priori. Cette expression est définie pour $a_{ij} > 0$, mais on se limitera à $a_{ij} \geqslant 1$, (seul cas utilisé en pratique). La loi a priori de P est donc $\prod_{i=1}^{r} f_i(p_i)$.

Tenant compte de (10) on peut maintenant écrire la densité conjointe a postériori, des observations

$$f(p_* | W) = K \left[\prod_{t=1}^{T} \prod_{j=1}^{r} \rho_j(t)^{n_j(t)} \right] \times \left[\prod_{i=1}^{r} \prod_{j=1}^{r-1} p_{ij}^{a_{ij}-1} \left(1 - \sum_{k=1}^{r-1} p_{ik}\right)^{a_i - \sum_{k=1}^{r-1} a_{ik} - 1} \right]$$

où K est une constante indépendante des paramètres.

La minimisation de - Log $f(p_* | W)$ fournit les estimateurs bayesiens de p_*. Cette minimisation s'interprète très simplement, comme on le verra au paragraphe suivant, dans une situation voisine.

II - Une interprétation bayesienne de la méthode de pénalisation dans l'estimation d'un vecteur de probabilités

1 - On envisage le cas où la méthode d'estimation utilisée (M.L. ou L.S.) risque de donner des estimateurs $\hat{p}$ n'appartenant pas à l'espace paramétrique

$$\Omega = \{p = (p_1, \ldots, p_n) \; ; \; p_i \geqslant 0 \quad i = 1, \ldots, n \text{ et } 1 - \sum_{i=1}^{n-1} p_i \geqslant 0\}$$

On est donc amené à faire intervenir explicitement les contraintes $\hat{p} \in \Omega$ dans la recherche des extremums.

2 - Estimateurs bayesiens

Hypothèse : le vecteur de probabilité p admet une loi a priori Beta-multidimensionnelle de paramètres $(a_1, \ldots, a_n)$ où $a_i > 1$ pour $i=1, \ldots, n$.

Remarque : La loi Beta-multidimensionnelle est définie pour $a_i > 0$ $(i=1, \ldots, n)$ mais si un a_i est inférieur à 1 la densité a postériori n'est plus bornée sur Ω. D'autre part la condition $a_i > 1$ pour tout i est nécessaire pour le calcul par pénalisation.

La densité conjointe a priori s'écrit :

$$B(p) = B(p_1, \ldots, p_n) \propto \prod_{i=1}^{n-1} p_i^{a_i - 1} \left(1 - \sum_{i=1}^{n-1} p_i\right)^{a_n - 1}$$

$L(X|p)$ étant la vraisemblance des observations, on peut choisir comme estimateur bayesien de p, une valeur modale $\hat{p}$ de la densité a posteriori. Nous admettons que cette valeur est unique.

Donc, en notant $f(p|X)$ la densité a postériori, on a :

$$f(p|X) \propto L(X|p)\ B(p)$$

et on cherche $\hat{p} \in \Omega$ tel que :

$$\min_{p \in \Omega} - \text{Log } f(p|X) = - \text{Log } L(X|\hat{p}) - \text{Log } B(\hat{p})$$

ce qui revient à chercher

$$\min_{p \in \Omega} - \text{Log } L(X|p) - \sum_{i=1}^{n-1} (a_i - 1) \text{ Log } p_i - (a_n - 1) \text{ Log}\left(1 - \sum_{i=1}^{n-1} p_i\right)$$

Cette expression n'est définie que sur Ω et tend vers $+\infty$ quand p se rapproche de la frontière. La recherche de son minimum se ramène à la recherche d'un minimum libre. C'est le principe des méthodes de pénalisation.

3 - Estimation M. L.

Si on recherche par une méthode de pénalisation, l'estimateur M. L. de p, minimum de $- \text{Log } L(X|p)$, $p \in \Omega$, on est amené à chercher une suite de minimums libres de la fonction $- \text{Log } L(X|p)$, pénalisée par une fonction des contraintes quand cette pénalisation tend vers 0. Cette pénalisation peut être l'expression :

$$+ P(\varepsilon;p) = + \left(\sum_{i=1}^{n-1} \varepsilon_i \text{ Log } p_i + \varepsilon_n \text{ Log}\left(1 - \sum_{j=1}^{n-1} p_j\right)\right)$$

où $\varepsilon_i > 0$ pour tout i. On calcule donc le minimum libre de

$$g(p;\varepsilon) = - \text{Log } L(X|p) - P(\varepsilon;p), \text{ où } \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$$

Soit $\hat{p}(\varepsilon)$ le point pour lequel $g(p;\varepsilon)$ est minimum.

Quand $\varepsilon \to 0$, $\hat{p}(\varepsilon) \to \hat{p}$, [4], $\hat{p}$ est l'estimateur M. L. de p.

Dans la solution numérique de ce problème, on arrête le calcul quand ε_i est petit mais strictement positif pour tout i, on n'obtient donc pas l'estimateur M. L. de p mais une valeur approchée qui dépend de ε : $p(\varepsilon)$. Si on se reporte à la discussion du numéro 2, on constate que l'approximation $\hat{p}(\varepsilon)$ de l'estimateur M. L. ainsi obtenu en minimisant $g(p;\varepsilon)$ pour $p \in \Omega$, est aussi solution du problème :

$$\min_{p\in\Omega} - \text{Log } L(X|p) - \sum_{i=1}^{n-1} (a_i - 1) \text{ Log } p_i - (a_n - 1) \text{ Log } (1 - \sum_{i=1}^{n-1} p_i)$$

Quand $a_i = \varepsilon_i + 1$ pour $i = 1,\ldots,n$.

Donc l'approximation $\hat{p}(\varepsilon)$ de l'estimateur M. L. $\hat{p}(0)$, est l'estimateur bayesien de p, obtenu en maximisant la densité a postériori, quand la densité a priori est une loi Beta-multidimensionnelle de paramètres : $(\varepsilon_1+1,\ldots,\varepsilon_n+1)$.

Ces remarques restent valables pour une large classe d'espaces paramétriques bornés. L'extension à un domaine non borné est plus délicate.

REFERENCES

[1] ANDERSON T.W. et GOODMAN L.A. (1957) - "Statistical Inference About Markov Chains", (The Ann. of Math. Stat. Vol. XXVIII, p. 89-110).

[2] ARAGON Y. (1971) - "Sur l'estimation des probabilités de transition d'une chaîne de Markov à partir des données marginales". (Thèse de 3ème cycle, Toulouse).

[3] DOUSSET F., CASAMITJANA R., GROBOILLOT J.L. & WARNIER DE WALLY A. (1967) - "Estimation d'une matrice de Markov" (Rev. Stat. Appl. Vol. 1, p. 87-94).

[4] FIACCO A.V. & MAC CORMICK G.P. (1968) - "Non Linear Programming : Sequential Unconstrained Minimization Technique" (Wiley).

[5] LEE T.C., JUDGE G.G. & CAIN R. (1967) - "A sampling of the Properties of Estimators of Transition Probabilities" (Agricultural Economics, Res. Report n° 87, University of Illinois).

[6] LEE T.C., JUDGE G.G. & TAKAYAMA T. (1965) - "On Estimating the Transition Probabilities of a Markov Process" (Journ. of Farm Econ., Vol. 47, p. 742-62).

[7] LEE T.C., JUDGE G.G. & ZELLNER A. (1968) - "Maximum Likelihood and Bayesian Estimation of Transition Probabilities" (J. Am. Stat. Ass., Vol. 63, p. 1162-1179).

[8] MC GUIRE T. (1969) - "More on Least Squares Estimation of the Transition Matrix in a Stationary First-Order Markov Process from sample Proportions Data" (Psychometrika, Vol. 34, n° 3, p. 335-346).

[9] MADANSKY A. (1959) - "Least Squares Estimation in Finite Markov Processes" (Psychometrika, Vol. XXIV, p. 137-44).

[10] MARQUARDT D.W. (1963) - "An Algorithm for Least-Squares Estimation of Non Linear Parameters" (Journ. Soc. Ind. Appl. Math., Vol. II, n° 2, p. 431-441).

[11] TELSER, LESTER G. (1966) - "Least Squares Estimates of Transition Probabilities" (Measurement in Economics, Stanford : Stanford University Press, p. 270-93).

[12] ZELLNER A. (1962) - "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias" (J. Am. Stat. Ass., Vol. 57, p. 348-68).