

ALGORITHME DE SÉLECTION SOUS CONTRAINTES D'UN SOUS-ENSEMBLE D'ÉLÉMENTS HOMOGÈNES Marcel G. Dagenais et Michel Dagenais

1. Introduction

Cet article a pour but de présenter un algorithme pour choisir un certain nombre d'éléments appartenant à des groupes déterminés, de façon à ce que les éléments retenus aient une variance minimale.

Une version simple du problème consiste à supposer que l'on a N groupes constitués chacun de deux nombres réels et qu'il faut choisir un élément par groupe, de façon à ce que la variance des éléments retenus soit minimale. La solution de ce problème par pur tâtonnement requerrait 2^N essais. On peut démontrer que la procédure que nous proposons nécessitera au plus $N + 1$ essais. En pratique, il faudra en général un nombre d'essais beaucoup plus petit.

Notre algorithme se généralise aussi au cas où chaque groupe i ($i = 1, \dots, N$) contient M_i éléments et où il faut choisir K_i éléments par groupe, de façon à ce que la variance globale des K éléments retenus (où $K = \sum_{i=1}^N K_i$) soit minimale.

La version simple du problème peut s'exprimer de la façon suivante:

$$\text{minimiser } \phi = \sum_{i=1}^N [Q_{1i}Z_i + Q_{2i}(1 - Z_i) - \bar{Q}]^2 \quad (1)$$

par rapport à $Z_1, \dots, Z_n$, où $\bar{Q} = \sum_{i=1}^N [Q_{1i}Z_i + Q_{2i}(1 - Z_i)]/N$,

(Q_{1i}, Q_{2i}) sont les valeurs des deux éléments du groupe i ,

$Z_i = 1$ si le premier élément du groupe i est choisi,

$Z_i = 0$ autrement, i.e., si le second élément du groupe i est retenu.

On peut également donner une interprétation statistique à notre problème, si l'on suppose qu'on dispose de N observations tirées au hasard d'une population de moyenne inconnue $\tilde{Q}$ et qu'on a deux mesures pour chaque observation, dont l'une est exacte et l'autre erronée; cependant, on ne sait pas laquelle des deux mesures est correcte. On peut alors obtenir un estimateur des moindres carrés de la moyenne de la façon suivante:

$$\text{minimiser } \phi^* = \sum_{i=1}^N [Q_{1i}Z_i + Q_{2i}(1 - Z_i) - \tilde{Q}]^2 \quad (2)$$

par rapport à $\tilde{Q}$, $Z_1, \dots, Z_N$.

Il est évident que pour tout sous-ensemble de Q_{ji} retenus ($j = 1$ ou 2 ; $i = 1, \dots, N$), la valeur de $\tilde{Q}$ qui minimise ϕ^* est $\tilde{Q} = \bar{Q}$. Par conséquent, cette interprétation alternative se réduit au problème déjà décrit dans l'équation (1).

On trouvera une application découlant de cette interprétation statistique dans Dagenais (1980). Le modèle considéré dans cet article est un modèle économique de "déséquilibre", dans lequel la rigidité des prix empêche le marché d'atteindre le point d'équilibre entre l'offre et la demande, à l'intérieur de la période considérée. En conséquence, la quantité de biens effectivement échangée pendant la période correspond *soit* à un point sur la courbe d'offre, *soit* à un point sur la courbe de demande.

Sous une forme simplifiée, les courbes d'offre et de demande peuvent s'écrire de la façon suivante:

$$D_i = \tilde{Q} + X_{di}\beta + u_{1i} \quad (i = 1, \dots, N)$$

$$S_i = \tilde{Q} + X_{si}\psi + u_{2i}$$

où D = quantité demandée, S = quantité offerte; X_{di} , X_{si} = vecteurs de variables explicatives; $\tilde{Q}$, $\check{Q}$ = termes constants; β , ψ = vecteurs de paramètres; u_{1t} , u_{2t} = erreurs résiduelles; i désigne la période de temps, D et S ne sont pas observés. On observe Q , la quantité échangée; celle-ci peut être égale soit à D ou à S , selon que le prix courant est supérieur ou inférieur au prix d'équilibre (inconnu). Le prix courant est une des variables explicatives contenues dans X_{si} et X_{di} .

Dans une première phase de la solution, en utilisant l'ensemble du système d'équations présenté dans l'article, l'auteur suggère une méthode qui permet d'estimer les vecteurs β et ψ ainsi que de la différence (d) entre $\tilde{Q}$ et $\check{Q}$. Ensuite, l'algorithme qui est présenté ici est utilisé pour obtenir un estimateur de $\tilde{Q}$. Si $Q_i = D_i$, on a $Q_{1i} = Q_i - X_{di}\beta = \tilde{Q} + u_{1i}$; si $Q_i = S_i$, on a $Q_{2i} = Q_i - X_{si}\psi - d = \check{Q} + u_{2i}$. La solution obtenue par notre algorithme minimise la somme des carrés des erreurs résiduelles, puisqu'on peut écrire

$$\phi^* = \sum_{i=1}^N [u_{1i} + u_{2i}(1 - Z_i)]^2.$$

2. La solution

La solution proposée découle de la proposition suivante qui montre que lorsque la valeur $\bar{Q}$ associée au minimum de ϕ est connue, il est facile d'identifier les éléments formant la solution.

PROPOSITION. Soit $\bar{Q}$ la valeur moyenne associée au choix qui minimise ϕ . Alors pour chaque groupe i , l'élément Q_{1i} (Q_{2i}) doit être l'élément retenu si et seulement si

$$|Q_{1i} - \bar{Q}| < |Q_{2i} - \bar{Q}|, \quad (|Q_{1i} - \bar{Q}| > |Q_{2i} - \bar{Q}|)$$

DÉMONSTRATION. Il est clair que la procédure proposée minimise ϕ . Montrons que si ϕ est minimale et que $\bar{Q}$ est la moyenne correspondante, alors la sélection fut faite suivant la règle proposée.

Supposons que pour un groupe i^* , $|Q_{1i^*} - \bar{Q}| < |Q_{2i^*} - \bar{Q}|$ et que Q_{2i^*} soit choisi, on aurait alors

$$\phi(Q_{2i^*}, \bar{Q}) > \phi(Q_{2i^*}, \bar{Q}) - (Q_{2i^*} - \bar{Q})^2 + (Q_{1i^*} - \bar{Q})^2 > \phi(Q_{1i^*}, Q^*) \quad (3)$$

où $\phi(Q_{2i^*}, \bar{Q})$ est la valeur de ϕ associée à la solution contenant Q_{2i^*} et pour laquelle la moyenne des éléments retenus est $\bar{Q}$, $\phi(Q_{1i^*}, Q^*)$ est la valeur obtenue pour ϕ après avoir remplacé Q_{2i^*} par Q_{1i^*} et $\bar{Q}$ par la nouvelle moyenne Q^* . L'inégalité de droite en (3) s'explique par le fait que dans le membre de droite on a une somme de carrés de déviations de certains éléments par rapport à leur moyenne, tandis que dans le membre de gauche, on a une somme de carrés des déviations des mêmes éléments par rapport à un autre point $(\bar{Q} \neq Q^*)$.

De l'équation (3), on déduit que $\phi(Q_{2i^*}, \bar{Q})$ ne peut être une solution puisque $\phi(Q_{2i^*}, \bar{Q}) > \phi(Q_{1i^*}, Q^*)$.

Q.E.D.

REMARQUE. Nous n'avons pas considéré le cas où $|Q_{1i^*} - \bar{Q}| = |Q_{2i^*} - \bar{Q}|$ car si $\bar{Q}$ est la moyenne des éléments associés à la solution de min. ϕ et que $|Q_{1i^*} - \bar{Q}| = |Q_{2i^*} - \bar{Q}|$, alors on doit nécessairement se trouver dans la situation triviale où $Q_{1i^*} = Q_{2i^*}$. En effet, si on avait $|Q_{1i^*} - \bar{Q}| = |Q_{2i^*} - \bar{Q}|$ et $Q_{1i^*} \neq Q_{2i^*}$, il s'en suivrait nécessairement que la solution considérée est inacceptable. En effet, supposons qu'un tel cas se présente et que la valeur contenue dans la solution, pour le groupe i^* , soit Q_{2i^*} , on aurait:

$$\phi(Q_{2i^*}, \bar{Q}) = \phi(Q_{2i^*}, \bar{Q}) - (Q_{2i^*} - \bar{Q})^2 + (Q_{1i^*} - \bar{Q})^2 > \phi(Q_{1i^*}, Q^*) \quad (4)$$

On voit clairement que $\phi(Q_{2i^*}, \bar{Q})$ ne pourrait être une solution. Le même raisonnement vaudrait évidemment si la valeur de Q contenue dans la solution était Q_{1i^*} au lieu de Q_{2i^*} .

Une fois établie, la proposition que nous venons d'énoncer, la première étape de la solution de notre problème consiste à calculer le point milieu de chaque groupe: $Q_{mi} = (Q_{1i} + Q_{2i})/2$. On doit ensuite ordonner les points milieux

dans l'ordre croissant. Les points milieux peuvent alors servir à diviser la droite réelle correspondant à toutes les valeurs possibles de $\bar{Q}$, en $N + 1$ segments associés aux $N + 1$ solutions possibles de $\min. \phi$. En effet, pour toutes les valeurs de $\bar{Q}$ contenues dans un segment donné, les Q_{ji} ($j = 1, 2 ; i = 1, \dots, N$) à retenir pour minimiser ϕ sont nécessairement les mêmes, en raison de l'affirmation contenue dans la proposition ci-dessus.

Afin d'être plus explicites, considérons l'exemple donné dans le Tableau 1 où $N = 5$ et où les groupes ont été ordonnés selon les valeurs croissantes des Q_{mi} . Sans perte de généralité, les valeurs des Q_{ji} , dans chaque groupe, ont été numérotées de façon à ce que $Q_{1i} \leq Q_{2i}$.

Q_{1i}	Q_{2i}	Q_{mi}
-30	14	- 8
-40	38	- 1
1	3	2
0	24	12
-15	217	101

Tableau 1

Les N valeurs inscrites dans la colonne des Q_{mi} divisent la ligne correspondant à $\bar{Q}$ en $N + 1$ segments, comme suit:

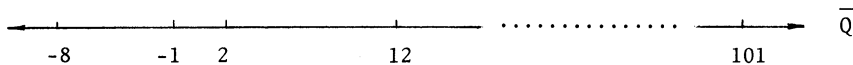


Figure 1

Par exemple, s'il existe une solution pour laquelle la valeur de $\bar{Q}$ est située dans l'intervalle $]2, 12[$, cette solution devrait contenir l'élément Q_{2i} pour tous les groupes où $Q_{mi} < 12$, et l'élément Q_{1i} pour tous les groupes où $Q_{mi} \geq 12$. Donc, la solution associée à ce segment contiendrait les éléments suivants: (14, 38, 3, 0, -15), ce qui donne $\bar{Q} = 8$.

On doit noter que très souvent il n'y aura aucune solution associée à un segment donné. Par exemple, si $\bar{Q}$ devait se trouver dans l'intervalle $] -1, 2[$, la seule solution admissible devrait alors contenir les éléments $(14, 38, 1, 0, -15)$, mais la moyenne de ces cinq valeurs est $\bar{Q} = 7.672$, ce qui est contradictoire.

3. L'algorithme

L'algorithme que nous proposons pour minimiser ϕ contient six étapes.

- 1) Dans chaque groupe, numéroter les éléments de façon à ce que l'on ait $Q_{1i} \leq Q_{2i}$.
- 2) Calculer le point milieu Q_{mi} ($i = 1, \dots, N$) de chaque groupe et re-numéroter les groupes selon les valeurs ascendantes des Q_{mi} . Si deux ou plusieurs groupes ont le même point milieu, il suffit d'attribuer à ces groupes des numéros consécutifs, de façon arbitraire.
- 3) Calculer $\bar{Q}_L = \sum_{i=1}^N Q_{1i}/N$ et $\bar{Q}_U = \sum_{i=1}^N Q_{2i}/N$. Il existe nécessairement un élément ℓ de $\{0, 1, \dots, N-1\}$ tel que $Q_{m\ell} < \bar{Q}_L \leq Q_{m(\ell+1)}$ ($Q_{m,0} = -\infty$ par définition). De la même façon, il existe nécessairement un élément s de $\{1, \dots, N\}$ tel que $Q_{ms} \leq \bar{Q}_U < Q_{m(s+1)}$ ($Q_{m(N+1)} = \infty$ par définition).
- 4) Diviser le segment $[\bar{Q}_L, \bar{Q}_U]$ de la droite réelle en r intervalles (où $r = s - \ell + 1$), en utilisant les points milieux Q_{mi} contenus dans ce segment.

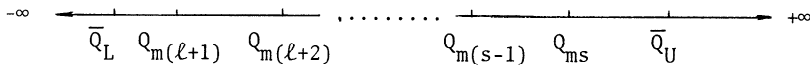


Figure 2

On aura évidemment $r \leq N+1$ et chacun des r intervalles contiendra au plus une solution.

- 5) Identifier les intervalles qui peuvent contenir une solution.
- 6) Calculer la valeur de ϕ pour toutes les solutions possibles et identifier la ou les solutions de min. ϕ .

Notez que, ainsi que nous l'avons mentionné à l'étape 2, les groupes qui ont le même point milieu peuvent être numérotés dans un ordre consécutif *arbitraire* car, comme nous l'avons démontré, la valeur de $\bar{Q}$ qui minimise ϕ ne peut jamais coïncider avec un point milieu; en conséquence, dans la solution finale, on sera amené, pour tous les groupes qui ont un même point milieu, à sélectionner soit tous les Q_1 ou tous les Q_2 .

Par exemple, en appliquant notre algorithme à l'exemple du Tableau 1, la Figure 2 de l'étape 4 devient:

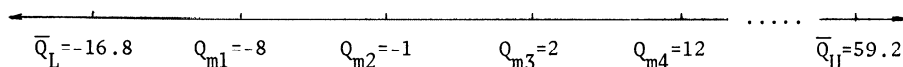


Figure 3

Ensuite, pour effectuer l'étape 5, on peut construire le tableau suivant:

Cas	Intervalles	Eléments contenus dans la solution	Valeur correspondante de $\bar{Q}$	$\bar{Q}$ est-il contenu dans l'intervalle correspondant
1	$[-16.8, -8[$	-30, -40, 1, 0, -15	-16.8	oui
2	$] -8, -1[$	14, -40, 1, 0, -15	- 8.0	non ¹
3	$] -1, 3[$	14, 38, 1, 0, -15	7.6	non
4	$] 2, 12[$	14, 38, 3, 0, -15	8.0	oui
5	$] 12, 59.2]$	14, 38, 3, 24, -15	12.8	oui

Tableau 2

Les solutions possibles sont celles qui correspondent aux cas 1, 4 et 5 . Les valeurs de ϕ associées à ces trois cas sont respectivement: 1 314.8 , 1 554 , 1 630.76 . Par conséquent, le minimum correspond au cas 1.

¹ Notez que dans ce cas, la valeur de ϕ est égale à un point milieu; cette valeur ne peut donc correspondre à une solution, comme on l'a démontré plus haut.

Dans cet exemple très simple, on constate que le nombre de cas pour lequel on a dû calculer ϕ s'élève à $3 < N+1 = 6$.

4. Expériences numériques

L'algorithme a été programmé sur un IBM 360-70. Les expériences ont porté sur deux exemples contenant chacun 1 200 groupes. Les valeurs des Q furent tirées au hasard à partir d'une loi normale centrée réduite. Dans le premier exemple, on a dû calculer ϕ seulement 35 fois, alors que pour le second exemple, il a fallu calculer ϕ , 41 fois. Le temps CPU requis pour solutionner les deux exemples, y compris la génération des nombres au hasard, a été de 3.92 secondes.

5. Généralisation

On peut généraliser la procédure décrite plus haut au cas où chaque groupe contient M_i éléments ($M_i \geq 1$; $i = 1, \dots, N$) parmi lesquels on doit en retenir K_i ($0 \leq K_i \leq M_i$). Il reste vrai de dire que si $\bar{Q}$ est connu, les K_i éléments du groupe i à retenir sont nécessairement ceux qui sont le plus près de $\bar{Q}$. En conséquence, pour traiter le cas général, il suffit de modifier notre algorithme de la façon suivante:

1) Dans chaque groupe, numéroter les M_i valeurs dans l'ordre croissant. Si deux ou plusieurs valeurs sont égales, il suffit de les numéroter consécutivement de façon arbitraire.

2) Pour chaque groupe, il faut ensuite localiser les $M_i - K_i$ points de division auxquels correspondent un nombre maximum de $M_i - K_i + 1$ sous-ensembles d'éléments pouvant être associés à des solutions de min. ϕ . On doit noter, en effet, que toute solution devra nécessairement inclure K_i éléments consécutifs dans le groupe i et que le nombre de sous-ensembles de K_i éléments consécutifs contenus dans le groupe i est égal à $M_i - K_i + 1$. Les $M_i - K_i$ points de division Q_{mji} ($j = 1, \dots, M_i - K_i$) sont obtenus en utilisant la formule suivante:

$$\phi_{mji} = (Q_{ji} + Q_{(j+K_i)i})/2 . \quad (5)$$

3) Calculer la moyenne $(\bar{Q}_L)$ pour l'ensemble des NK_i éléments contenant les K_i plus petites valeurs de chaque groupe et calculer aussi la moyenne $(\bar{Q}_U)$ associée aux K_i plus grandes valeurs de chaque groupe.

4) Localiser sur la droite réelle, les valeurs de $\bar{Q}_L$ et $\bar{Q}_U$ ainsi que tous les points de division identifiés à l'étape 2) qui sont situés entre $\bar{Q}_L$ et $\bar{Q}_U$, i.e. localiser tous les Q_{mji} ($j = 1, \dots, M_i - K_i$; $i = 1, \dots, N$) tels que $\bar{Q}_L < Q_{mji} < \bar{Q}_U$. Le segment $[\bar{Q}_L, \bar{Q}_U]$ sera alors par le fait même divisé en, au plus, $\sum_{i=1}^N M_i - \sum_{i=1}^N K_i + 1$ intervalles. Chacun de ces intervalles pourra correspondre à au plus une solution.

5) Identifier les intervalles qui peuvent effectivement contenir une solution.

6) Calculer la valeur de ϕ pour chacune des solutions possibles et identifier la ou les solutions de min. ϕ .

Cette recherche a été effectuée avec la collaboration de M. Tran Cong Liem. Le projet a été subventionné principalement par le Conseil des arts du Canada. Il a aussi bénéficié en partie d'un octroi provenant du ministère de l'Éducation du Québec. Nous tenons enfin à remercier un des arbitres dont les suggestions ont grandement contribué à améliorer la présentation du texte.

6. Références

DAGENAIS, Marcel G., *Specification and estimation of a dynamic disequilibrium model*, Economics Letters, Vol. 5, No 4, 1980, 323-328.