RAIRO. RECHERCHE OPÉRATIONNELLE

VIDAL COHEN

# Application de l'optimisation en nombres entiers à un problème d'« arrondissage »

*RAIRO. Recherche opérationnelle*, tome 16, n° 4 (1982), p. 365-377

<http://www.numdam.org/item?id=RO_1982__16_4_365_0>

R.A.I.R.O. Recherche opérationnelle/Operations Research
(vol. 16, n° 4, novembre 1982, p. 365 à 376)

# APPLICATION DE L'OPTIMISATION EN NOMBRES ENTIERS A UN PROBLÈME D'« ARRONDISSAGE » (*)

par Vidal COHEN (1)

Résumé. — *L'approximation en nombres entiers (ou « arrondissage ») d'un ensemble de réels ne respecte pas, en général, les opérations définies sur celui-ci. Nous proposons ici une procédure de moindres carrés corrigeant cet inconvénient et l'appliquons à des listes et à des tableaux pour des fonctions particulières dont l'une au moins conduit à un résultat assez inattendu.*

Mots clés : Arrondissage de listes et de tableaux; approximation en nombres entiers.

Abstract. — *The integer approximation (or rounding off) of a set of real numbers does not generally respect the operations defined upon it. A least-squares procedure is presented here to tackle this drawback. We apply it to lists and tables for some elementary functions. One example leads to a rather unexpected result.*

Keywords: Rounding off of lists and tables; Integer approximation.

## 1. INTRODUCTION

Lorsque les éléments réels d'une liste ou d'un tableau sont arrondis à des entiers, pour des raisons diverses souvent liées à l'imprécision des données, le problème suivant se pose : à la suite de certaines opérations (addition, calcul de marges, ...), les arrondis des résultats diffèrent en général de ce qu'auraient donné les mêmes opérations effectuées sur les éléments préalablement arrondis. Ce problème, abordé dans deux publications récentes [1, 4] est ici traité dans une perspective différente : celle de fournir une procédure d'approximation automatisable, applicable dans les cas qu'elles mentionnent où elle élimine l'inconvénient signalé.

(*) Reçu juillet 1981.

(1) Université de Paris IX-Dauphine, place du Maréchal-de-Lattre-de-Tassigny, 75116 Paris.

R.A.I.R.O. Recherche opérationnelle/Operations Research, 0399-0559/1982/365/$ 5.00

## 2. ARRONDISSAGE « OPTIMAL » DANS UN CAS DE LISTE

### 2.1. Présentation du problème

Soient un ensemble fini :

$$E = \{ x_i / x_i \in \mathbb{R};\ i = 1 \ldots n \}$$

et :

$$f(E) = f(x_1, \ldots, x_i, \ldots, x_n) = \sum_{i=1}^{n} x_i.$$

Si $g : \mathbb{R} \mapsto \mathbb{Z}$ associe à tout réel $a$ l'entier $z$ (pour $a$ demi-entier, l'un, spécifié, des deux entiers) le plus proche de $a$ au sens de l'écart $|z-a|$, il est clair qu'en général :

$$f(g(x_i)/i = 1 \ldots n) \neq g(f(E)),$$

*Exemple :* Si :

$$E = \{x_1 = 2{,}60;\ x_2 = -1{,}20;\ x_3 = 1{,}55\},$$

$$f(E) = 2{,}95 \qquad \text{d'où} \quad g(f(E)) = 3,$$

tandis que :

$$f(g(x_i)/i = 1,\ 2,\ 3) = f(3;\ -1;\ 2) = 4.$$

### 2.2. Arrondissage aux moindres carrés respectant l'addition

Nous chercherons un élément de $\mathbb{Z}^n$, soit $\{ \tilde{g}(x_i)/i = 1, \ldots, n \}$, tel que :

(I)
$$\sum_{i=1}^{n} [\tilde{g}(x_i) - x_i]^2 = \operatorname*{Min}_{u_i \in \mathbb{Z}} \sum_{i=1}^{n} (u_i - x_i)^2,$$

les entiers $u_i$ étant soumis à la contrainte :

$$f(u_1, \ldots, u_i, \ldots, u_n) = \sum_{i=1}^{n} u_i = g(f(x_1, \ldots, x_i, \ldots, x_n)).$$

La procédure dont nous justifierons ultérieurement le choix résulte des lemmes ci-après.

LEMME 1 : *Le n-uplet* $E^0 = \{ y_i / i = 1 \ldots n \} \in \mathbb{R}^n$ *le plus proche de* $E = \{ x_i / i = 1 \ldots n \}$ *au sens des moindres carrés, sous la contrainte :*

$$f(E^0)\,(= \sum_i y_i) = g(f(E)) \qquad (= g(\sum_i x_i))$$

*est :*

$$E^0 = \left\{ y_i = x_i + A / A = \frac{g(\sum_i x_i) - \sum_i x_i}{n};\ i = 1 \ldots n \right\}. \tag{1}$$

Cette relation (1) résulte des conditions de Lagrange, nécessaires et suffisantes pour ce minimum d'une fonction convexe sous contrainte affine.

Si $(\forall i)$, $y_i \in \mathbb{Z}$, ce sont les arrondis aux moindres carrés cherchés. Sinon :

LEMME 2 : *Les arrondis* $\tilde{g}(x_i)$ *cherchés sont :* $\tilde{g}(x_i) = y_i + z_i$ $(i = 1 \ldots n)$ *où les* $y_i$ *sont donnés par* (1) *tandis que les* $z_i$ *doivent vérifier :*

$$\text{(II)} \qquad \left\{ \begin{array}{l} \text{Min} \sum_{i=1}^{n} z_i^2, \\ \text{sous les contraintes} \left\{ \begin{array}{l} \sum_i z_i = 0, \\ (\forall i) \quad (y_i + z_i) \in \mathbb{Z}. \end{array} \right. \end{array} \right.$$

Ces contraintes résultent en effet des qualités attendues des $\tilde{g}(x_i)$, à savoir :

$$\sum_i \tilde{g}(x_i) = g(\sum_i x_i) = \sum_i y_i \qquad \text{d'où} \quad \sum_i z_i = 0$$

et :

$$(\forall i), \quad \tilde{g}(x_i) \in \mathbb{Z},$$

tandis que l'exigence de minimisation de $\sum_i z_i^2$ découle de la relation :

$$\sum_i (y_i + z_i - x_i)^2 = \sum_i z_i^2 + 2 \sum_i z_i (y_i - x_i) + \sum_i (y_i - x_i)^2$$

$$= \sum_i z_i^2 + 2A \sum_i z_i + nA^2 = \sum_i z_i^2 + nA^2.$$

D'où :

$$\operatorname{Min} \sum_i (\tilde{g}(x_i) - x_i)^2 \quad \Leftrightarrow \quad \operatorname{Min} \sum_i z_i^2,$$

si les $\tilde{g}(x_i)$ sont solutions de (I).

Introduisant alors pour chaque $y_i$ sa partie entière $[y_i]$ (plus grand entier rationnel inférieur à $y_i$) et sa partie décimale $r_i (\geqq 0)$ soit :

$$y_i = [y_i] + r_i,$$

il est clair que les $z_i^*$ solutions de (II) doivent vérifier :

$$z_i^* = -r_i + Z_i \qquad \text{avec} \quad (\forall i), \quad Z_i \in \mathbb{Z}. \tag{2}$$

En fait, plus précisément :

LEMME 3 : *Les $Z_i$ associés par* (2) *aux $z_i^*$ solutions de* (II) *ne peuvent valoir que* 0 *ou* 1.

En effet, s'il existait un $z_k^*$ tel que :

$$Z_k = z_k^* + r_k < 0 \quad (\text{resp. } > 1),$$

il existerait nécessairement, puisque $\sum_i z_i^* = 0$ et que $(\forall i)$, $0 \leqq r_i < 1$, un $z_j^*$ au moins tel que $Z_j \geqq 1$ (resp. $\leqq 0$). Par suite, en remplaçant : $z_k^*$ par $z_k^* + 1$ et $z_j^*$ par $z_j^* - 1$ (resp. $z_k^*$ par $z_k^* - 1$ et $z_j^*$ par $z_j^* + 1$) le minimum $\sum_i z_i^{*2}$ se trouverait abaissé, ce qui est absurde (car par exemple : pour $Z_k < 0$, donc $Z_k \leqq -1$, et $Z_j \geqq 1$, on aurait :

$$(-r_k + Z_k + 1)^2 + (-r_j + Z_j - 1)^2 < (-r_k + Z_k)^2 + (-r_j + Z_j)^2,$$

puisque :

$$1^2 + 1^2 + 2(-r_k + Z_k) - 2(-r_j + Z_j) = 2(1 - r_k + Z_k + r_j - Z_j)$$
$$\leqq 2(-r_k + r_j - 1) < 0,$$

tandis qu'un raisonnement analogue est possible pour un $Z_k > 1$).

*Un mode de détermination des $Z_i$* en résulte :

$$\sum_i Z_i = \sum_i z_i^* + \sum_i r_i = \sum_i r_i = \sum_i y_i - \sum_i [y_i] = g\left(\sum_i x_i\right) - \sum_i [y_i] = K \in \mathbb{N}$$

$K$ étant ainsi un *entier naturel connu.*

Les $Z_i$ seront donc déterminés comme suit : il faudra choisir :

$Z_i = 1$ pour $K$ valeurs de l'indice $i$;

$Z_i = 0$ pour les $(n-K)$ valeurs restantes de $i$.

Mais :

$$\sum_i z_i^2 = \sum_i (-r_i + Z_i)^2 = \sum_i r_i^2 + \sum_i Z_i^2 - 2\sum_i r_i Z_i$$

$$= \sum_i r_i^2 + \sum_i Z_i - 2\sum_i r_i Z_i \qquad (\text{car } (\forall i),\ Z_i^2 = Z_i)$$

$$= \sum_i r_i^2 + K - 2\sum_i r_i Z_i.$$

D'où :

$$\text{Min} \sum_i z_i^2 \quad \Leftrightarrow \quad \text{Max} \sum_i r_i Z_i.$$

En conséquence :

LEMME 4 : *Il convient d'attribuer la valeur* 1 *à* $Z_i$ *pour les valeurs de l'indice* $i$ *associées aux* $K(=\sum_i r_i)$ *plus grandes valeurs des* $r_i$ *(une certaine latitude de choix subsistant en cas d'éventuels ex-aequo), les autres* $Z_i$ *prenant la valeur* 0.

De ce qui précède, il découle naturellement :

*Une procédure d'arrondissage de liste*

Si l'on arrondit la somme $\sum_i x_i$ des éléments du $n$-uplet réel $E = \{ x_i / i = 1 \ldots n \}$ à l'entier $S = g(\sum_i x_i)$, le $n$-uplet entier $\tilde{g}(E) = \{ u_i / i = 1 \ldots n \}$ respectant l'addition (i. e. tel que $\sum_i u_i = S$) et approchant $E$ au sens des moindres carrés s'obtient comme suit :

1° déterminer les $y_i$ :

$$y_i = x_i + \frac{1}{n}(S - \sum_i x_i);$$

2° si les $y_i$ ainsi calculés ne sont pas tous entiers, mettre en évidence leurs parties entières $[y_i]$ et décimales $r_i$ :

$$y_i = [y_i] + r_i;$$

3° ordonner les indices $i$ selon les valeurs décroissantes des $r_i$ et écrire les $y_i$ dans le même ordre;

4° les valeurs cherchées $u_i$ seront les arrondies *par excès* des $y_i$ pour les $K(=\sum_i r_i)$ premiers éléments de la liste ainsi constituée et les arrondies *par défaut* des $y_i$ pour les $n-K$ éléments suivants de cette liste (une certaine latitude de choix subsistant si la $K$-ième valeur de $r_i$ est réalisée pour plusieurs valeurs de l'indice $i$).

Dans le cas d'une liste, l'arrondissage aux moindres carrés s'effectuera toujours en arrondissant les parties décimales $r_i$ à l'un des entiers voisins 0 ou 1. Il en résulte que si la somme a été arrondie à l'entier le plus proche, chacun des termes ne pourra être arrondi qu'à l'un des deux entiers qui l'encadrent.

*Conséquence :* Au sens de l'entier le plus proche, l'arrondi d'une somme de réels $x_i$ est égal à la somme des arrondis des $x_i$ si et seulement si les $K$ plus grandes valeurs des parties décimales des $y_i$ associés sont comprises entre 0,5 et 1 ($K$ et les $y_i$ ayant été définis ci-dessus).

Retour à l'exemple mentionné en 2.1 :

$$\begin{cases} y_1 = 2{,}60 + \frac{1}{3}(3-2{,}95) = 2{,}6166\ldots, \\ y_2 = -1{,}20 + 0{,}0166\ldots, \\ y_3 = 1{,}55 + 0{,}0166\ldots, \end{cases} \qquad \text{d'où :} \quad \begin{cases} r_1 = 0{,}6166\ldots, \\ r_2 = 0{,}8166\ldots, \\ r_3 = 0{,}5666\ldots, \end{cases}$$

$$K = \sum_i r_i = 2.$$

On constate ici que $r_2 > r_1 > r_3$.

Dans les conditions précisées, les arrondis optimaux seront :

$$\tilde{g}(x_1) = 3; \qquad \tilde{g}(x_2) = -1; \qquad \tilde{g}(x_3) = 1$$

puisque $y_1$ et $y_2$ doivent être arrondis par excès et $y_3$ par défaut, ce qui d'ailleurs donne bien :

$$\sum_i \tilde{g}(x_i) = g\left(\sum_i x_i\right) = 3.$$

REMARQUE : La méthode d'arrondissage proposée ici peut sembler artificielle : en particulier, l'introduction des $y_i$ associés aux $x_i$ donnés initialement. Leur intérêt apparaît lorsque l'on décide d'arrondir la somme $\sum_i x_i$ non pas à l'entier le plus voisin, mais à un élément d'un sous-ensemble de

$\mathbb{Z}$ (par exemple à l'élément le plus proche du sous-anneau des multiples de 5 ou de 10). Le passage par les $y_i$ permet en fait, en bénéficiant du lemme 3, d'éviter (grâce à la relation $Z_i^2 = Z_i$) une optimisation quadratique.

Ajoutons que la transformation $\tilde{g}$ se trouve ainsi définie comme une application de $\mathbb{R}^n$ dans $\mathbb{Z}^n$et non de $\mathbb{R}$ dans $\mathbb{Z}$.

Nous allons maintenant fournir une présentation plus générale.

### 3. UNE MÉTHODE D'ARRONDISSAGE

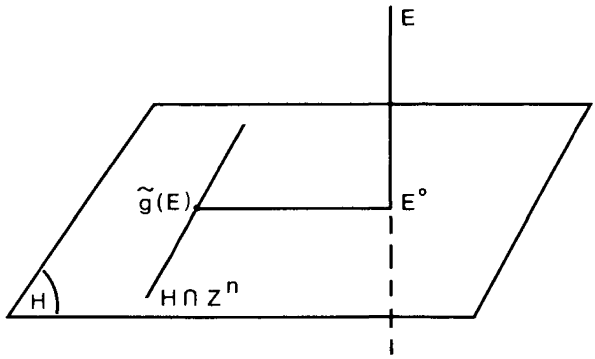
Généralisons quelque peu le problème précédent :

Soient un point $E \in \mathbb{R}^n$ (affine euclidien) et une application linéaire $f : \mathbb{R}^n \mapsto \mathbb{R}^k$ ($n$ et $k$ entiers naturels fixés; dans l'exemple ci-dessus : $k = 1$).

Si l'on note $g(f(E))$ l'arrondi dans $\mathbb{Z}^k$, selon une règle définie, de l'image $f(E) \in \mathbb{R}^k$, on envisagera dans $\mathbb{R}^n$ la variété affine (contre-image) :

$$H = f^{-1}\{g(f(E))\} = \{x / x \in \mathbb{R}^n ; f(x) = g(f(E))\},$$

puis la projection orthogonale $E^0$ de $E$ sur $H$ (l'orthogonalité étant comprise au sens de la forme bilinéaire associée à la distance des moindres carrés choisie dans $\mathbb{R}^n$ — ici la forme canonique).




L'arrondi $\tilde{g}(E)$ sera alors à chercher dans $H \cap \mathbb{Z}^n$, le plus près possible de $E^0$.

On reconnaît que $E^0$ est précisément l'ensemble $\{y_i / i = 1 \ldots n\}$ introduit plus haut.

Cependant, si, dans le cas où $E$ est une liste et $f(E)$ une somme, $\tilde{g}(E)$ est l'un des sommets du plus petit pavé de $\mathbb{Z}^n$ contenant $E^0$, il n'en est plus toujours ainsi dans le cas général : nous mentionnerons brièvement deux exemples.

*Exemple* 3.1. *Forme linéaire à coefficients rationnels*

Si :

$$f(E)=f(x_1, \ldots, x_i, \ldots, x_n)=\sum_i q_i x_i \qquad (q_i \in \mathbb{Q}),$$

on pourra également écrire :

$$f(x_1, \ldots, x_i, \ldots, x_n)=\frac{1}{q}\sum_i a_i x_i \qquad (q \in \mathbb{N};\ a_i \in \mathbb{Z}).$$

Arrondissant $f(x_1, \ldots, x_i, \ldots, x_n)$ à un entier $g(f(x_1, \ldots, x_n)) \in \mathbb{Z}$, il sera facile de déterminer comme plus haut les $y_i \in \mathbb{R}$ approchant au mieux (au sens des moindres carrés) les $x_i$ et vérifiant :

$$\frac{1}{q}\sum_i a_i y_i = g(f(x_1, \ldots, x_i, \ldots, x_n)).$$

Si les $y_i$ sont décomposés en leurs parties entières $[y_i]$ et décimales $r_i$ :

$$y_i=[y_i]+r_i,$$

il faut, comme précédemment, trouver les $z_i \in \mathbb{R}$ réalisant :

$$\text{(III)} \quad \begin{cases} \text{MIN} \sum_i z_i^2 \text{ sous les contraintes :} \\ (\forall i), \quad y_i+z_i \in \mathbb{Z} \qquad \text{et} \qquad \sum_i a_i z_i = 0, \end{cases}$$

soit, posant :

$$z_i=-r_i+Z_i \qquad (Z_i \in \mathbb{Z}),$$

$$\sum_i a_i Z_i = \sum_i a_i r_i = K \qquad (\in \mathbb{Z}). \tag{3}$$

Pour que ces conditions soient réalisables, il est nécessaire et suffisant, d'après le théorème de Bezout, que le PGCD des $a_i$ divise $K$, ce qui équivaut à la divisibilité de $qg[f(E)]$ par le PGCD des $a_i$.

S'il en est ainsi, il existe des valeurs $Z_i^0$ des $Z_i$ qui vérifient (3) et que l'on détermine facilement, en utilisant d'ailleurs la procédure servant habituellement à démontrer ce théorème. Alors, posant :

$$b_i=Z_i^0-r_i,$$

le problème précédent prend la forme équivalente :

$$\text{(III')} \quad \begin{cases} \text{Recherche des } Z_i^* \in \mathbb{Z} \text{ réalisant :} \\ \qquad \text{MIN} \sum_i (b_i+Z_i)^2 \text{ sous la contrainte :} \\ \qquad\qquad \sum_i a_i Z_i = 0. \end{cases} \tag{4}$$

La contrainte (4) présente sur (3) l'avantage d'une caractérisation des entiers $Z_i$ indépendante des valeurs attribuées aux $x_i$ et à $g[f(E)]$. Partant de $Z_i = 0 (\forall i)$, on améliore le critère en considérant les couples successifs $(Z_i, Z_{i'})$ (utilisation du théorème de Gauss), ce qui permet de borner le domaine d'exploration utile dans $\mathbb{Z}^n$, le théorème de Bezout permettant alors d'y énumérer les $n$-uplets "admissibles".

Notons que la quantité :

$$(\sum_i a_i b_i)^2 / \sum_i a_i^2 ,$$

constitue un minorant du minimum cherché en tant que solution du problème précédent relaxé de la contrainte d'intégrité sur les $Z_i$.

*Application numérique*

Soit :

$$x = (x_1, x_2, x_3, x_4) \qquad (\text{donc } n = 4)$$

et soit :

$$f(x) = 4x_1 - 5x_2 + 2x_3 + 8x_4 \qquad (\text{donc : } a_1 = 4;\ a_2 = -5;\ a_3 = 2;\ a_4 = 8).$$

Si par exemple :

$$x_1 = 1{,}4; \qquad x_2 = 2{,}6; \qquad x_3 = 3{,}7; \qquad x_4 = 1{,}3,$$

il vient : $f(x) = 10{,}4$.

Arrondissant cette valeur à l'entier voisin, nous obtenons : $g(f(x)) = 10$.

Ici :

$$y_i = x_i + a_i (10 - 10{,}4) / \sum_i a_i^2,$$

d'où :

$$y_1 = 1{,}385; \qquad y_2 = 2{,}618; \qquad y_3 = 3{,}692; \qquad y_4 = 1{,}271;$$

et leurs parties décimales :

$$r_1 = 0{,}385; \qquad r_2 = 0{,}618; \qquad r_3 = 0{,}692; \qquad r_4 = 0{,}271;$$

soit : $K = \sum_i a_i r_i = 2$.

Il est aisé d'obtenir des $Z_i^0 \in \mathbb{Z}$ tels que : $\sum_i a_i Z_i^0 = 2$ dès lors que le PGCD des $a_i$ (ici égal à 1) divise $K$ (ici égal à 2) :

$$Z_1^0 = 2; \qquad Z_2^0 = 2; \qquad Z_3^0 = -6; \qquad Z_4^0 = 2$$

(en effet : $4\times 2-5\times 2+2\times(-6)+8\times 2=2$). Nous ignorons volontairement ici la solution visiblement optimale $(0,0,1,0)$ pour indiquer comment elle sera trouvée en général.

Nous obtenons ainsi un premier système (non optimal) de valeurs arrondies :

$$g'(x_1)=y_1+Z_1^0-r_1=3; \qquad g'(x_2)=4; \qquad g'(x_3)=-3; \qquad g'(x_4)=3$$

et comme :

$$b_1=1{,}615; \qquad b_2=1{,}382; \qquad b_3=-6{,}692; \qquad b_4=1{,}729,$$

il s'agit de rechercher les $Z_i^* \in \mathbb{Z}$ réalisant :

$$\text{MIN}\{1{,}615+Z_1)^2+(1{,}382+Z_2)^2+(-6{,}692+Z_3)^2+(1{,}729+Z_4)^2\},$$

sous la contrainte :

$$4Z_1+(-5)Z_2+2Z_3+8Z_4=0.$$

L'exploration des éléments "admissibles" de $\mathbb{Z}^4$ (selon la procédure mentionnée plus haut) conduit ici à :

$$Z_1^*=-2; \qquad Z_2^*=-2; \qquad Z_3^*=7; \qquad Z_4^*=-2;$$

et pour les valeurs initialement adoptées :

$$x_1=1{,}4; \qquad x_2=2{,}6; \qquad x_3=3{,}7; \qquad x_4=1{,}3,$$

aux valeurs arrondies :

$$\tilde{g}(x_1)=1; \qquad \tilde{g}(x_2)=\underline{2}; \qquad \tilde{g}(x_3)=4; \qquad \tilde{g}(x_4)=1.$$

*Exemple* 3.2. *Tableau* $X=(x_{ij})$ *d'ordre* $(I\times J)$ *à éléments réels*

Supposons qu'ici :

$$f(X)=\{x_{i.}=\sum_j x_{ij}\,(i=1\ldots I);\ x_{.j}=\sum_i x_{ij}\,(j=1\ldots J)\}$$

et que nous ayons arrondi ces « marges » $x_{i.}$ et $x_{.j}$ à des entiers notés respectivement $y_{i.}$ et $y_{.j}$ et vérifiant :

$$\sum_i y_{i.}=\sum_j y_{.j} \qquad (=y_{..})$$

(nous pourrons, par exemple, avoir déterminé ceux-ci par la procédure exposée en 2, $y_{..}$ étant précisément l'entier arrondi du total général $x_{..}$ de $X$).

Poursuivant l'application de notre méthode, il conviendra alors de construire le tableau $Y=(y_{ij})$ de même format que $X$, admettant précisément pour marges $y_{i.}$ $(i=1\ldots I)$; $y_{.j}$ $(j=1\ldots J)$ et le plus proche de $X$ au sens des

moindres carrés. Notons que les $y_{ij}$ jouent exactement le même rôle que les $y_i$ dans le cas de listes. Un tel tableau $Y$ se calcule aisément :

$$y_{ij}=\frac{y_{i.}}{J}+\frac{y_{.j}}{I}-\frac{y_{..}}{IJ}+x_{ij}-\frac{x_{i.}}{J}-\frac{x_{.j}}{I}+\frac{x_{..}}{IJ},$$

une extension étant même possible au cas de tableaux à plus de deux entrées [2]. Précisons que si chaque tableau est représenté par un vecteur de l'espace euclidien $\mathbb{R}^{IJ}$, le vecteur $X-Y$ sera orthogonal à tout vecteur $z=(z_{ij})$ correspondant à un tableau à marges nulles (c'est-à-dire vérifiant :

$$(\forall i\in\{1\ldots I\}),\ z_{i.}=\sum_j z_{ij}=0;\ (\forall j\in\{1\ldots J\}),\ z_{.j}=\sum_i z_{ij}=0),$$

en ce sens que :

$$\sum_{i,j}(y_{ij}-x_{ij})\,z_{ij}=0.$$

Décomposant alors comme précédemment les $y_{ij}$ en leurs parties entières $[y_{ij}]$ et leurs parties décimales $r_{ij}$ :

$$y_{ij}=[y_{ij}]+r_{ij},$$

il conviendra, après avoir posé :

$$z_{ij}=-r_{ij}+Z_{ij},$$

de rechercher les $Z_{ij}^*\in\mathbb{Z}$ réalisant :

(IV)
$$\text{MIN}\sum_{i,j}(Z_{ij}-r_{ij})^2,$$

sous les contraintes :

$$(\forall i\in\{1\ldots I\}),\quad Z_{i.}=\sum_j Z_{ij}=r_{i.},$$

$$(\forall j\in\{1\ldots J\}),\quad Z_{.j}=\sum_i Z_{ij}=r_{.j},$$

les $r_{i.}$ et $r_{.j}$ étant les entiers naturels constituant les marges du tableau $r=(r_{ij})$.

REMARQUE : Pour éviter d'arrondir les $y_{ij}$ à des entiers ne les encadrant pas, on pourrait imposer aux $Z_{ij}^*$ d'être égaux à 0 ou 1; on peut d'ailleurs montrer,

en utilisant la condition (C) de Fréchat [3], qu'un tel tableau, soumis de plus aux contraintes de marges, existe. Mais l'optimum peut se trouver amélioré, comme nous le verrons plus loin sur un exemple, si les $Z_{ij}^*$ sont des entiers pouvant prendre des valeurs autres que 0 ou 1.

*Détermination des $Z_{ij}^*$*

On pourra d'abord traiter les $I$ lignes du tableau $r$ comme dans le cas de liste. Si les colonnes sont toutes équilibrées, l'optimum est atteint avec $Z_{ij}^* \in \{0; 1\}$. Sinon, il conviendra de transférer des 1 des colonnes « excédentaires » (i. e. $Z_{.j} > r_{.j}$) vers les colonnes « déficitaires » (i. e. $Z_{.j} < r_{.j}$), mais en accroissant au minimum l'écart quadratique de $Z$ à $r$.

Ce qui est assez inattendu, c'est que dans des cas particuliers (que l'on peut caractériser) il peut y avoir avantage à effectuer de tels transferts jusqu'à faire apparaître certains $Z_{ij}^*$ égaux à $-1$ ou à 2 : ainsi, au sens des moindres carrés :

$$Y = \begin{array}{|ccccc|c} 0{,}25 & 0{,}25 & 0{,}25 & \underline{0{,}98} & 0{,}27 & 2 \\ 0{,}70 & 0{,}10 & 0{,}10 & 0{,}40 & 0{,}70 & 2 \\ 0{,}03 & 0{,}60 & 0{,}05 & 0{,}30 & 0{,}02 & 1 \\ 0{,}02 & 0{,}05 & 0{,}60 & 0{,}32 & 0{,}01 & 1 \\ \hline 1 & 1 & 1 & 2 & 1 & 6 \end{array}$$

sera arrondi au mieux à :

$$Z^* = \begin{array}{|ccccc|c} 0 & 0 & 0 & \underline{2} & 0 & 2 \\ 1 & 0 & 0 & 0 & 1 & 2 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ \hline 1 & 1 & 1 & 2 & 1 & 6 \end{array} .$$

On voit qu'un élément de $Y$ (égal à 0,98) se trouve arrondi à 2, entier différent des deux qui l'encadrent.

## 4. CONCLUSION

La méthode proposée est applicable dans des cas assez généraux. Mentionnons aussi la situation suivante rencontrée en Statistique (par exemple en analyse de données sur variables qualitatives) : si des modalités sont codées en variables indicatrices (0; 1), il arrive que le modèle « fournisse » des résultats en valeurs décimales qui n'auraient de sens qu'arrondis à 0 ou 1 : notre méthode serait-elle alors de quelque secours, l'écart quadratique choisi s'interprétant comme un « coût » ?

## BIBLIOGRAPHIE

1. F. CHARTIER, *Note sur l'arrondissage automatique des termes d'une somme*, Annales de l'INSEE, n° 25, 1977, p. 139-151.
2. V. COHEN, Les tableaux à *n* entrées : rappel de quelques questions et solution d'un problème de caractérisation, *Cahier du LAMSADE* (Univ. de Paris-IX), n° 39, 1982.
3. M. FRECHET, *Sur les tableaux dont les marges et des bornes sont données*, Revue de l'Inst. Internat. de Stat., vol. 28, n° 1/2, 1960, p. 10-32.
4. P. THIONET, *Note sur l'« arrondissage »*, Annales de l'INSEE, n° 25, 1977.