Mathématiques et sciences humaines
Mathematics and social sciences

187 | Automne 2009
Journée 2007 de la Société Francophone de Classification

Quelques remarques sur la méthode d'ajustement de Mayer : lien avec les méthodes de classifications

Mayer's fitting method and its links to clustering methods

Antoine Falguerolles

Édition électronique
URL : http://journals.openedition.org/msh/11117
DOI : 10.4000/msh.11117
ISSN : 1950-6821

Éditeur
Centre d'analyse et de mathématique sociales de l'EHESS

Édition imprimée
Date de publication : 30 décembre 2009
Pagination : 43-58
ISSN : 0987-6936

Référence électronique
Antoine Falguerolles, « Quelques remarques sur la méthode d'ajustement de Mayer : lien avec les méthodes de classifications », *Mathématiques et sciences humaines* [En ligne], 187 | Automne 2009, mis en ligne le 15 décembre 2009, consulté le 19 avril 2019. URL : http://journals.openedition.org/msh/11117 ; DOI : 10.4000/msh.11117

Math. & Sci. hum. / Mathematics and Social Sciences (47e année, n° 187, *2009*(3), p. 43–58)

LA MÉTHODE D'AJUSTEMENT DE MAYER ET SES LIENS AVEC LES MÉTHODES DE CLASSIFICATION

Antoine de FALGUEROLLES[1]

RÉSUMÉ – *Le cas simple de l'ajustement d'une droite de régression par la méthode de Mayer, au programme de l'enseignement secondaire français il y a quelques années, avait été introduit comme un succédané de la méthode des moindres carrés. Il apparaît que la démarche qui était ainsi proposée aux élèves fournit un exemple élémentaire d'arbre de régression. Il apparaît aussi que, dans le cas général, c'est un problème de classification pour lequel l'algorithme des transferts de Régnier [1965] est particulièrement bien adapté quoique possiblement suboptimal. L'exemple célèbre d'ajustement, que Mayer traite en 1750 par une méthode novatrice et très générale, est revu à la lumière de méthodes statistiques contemporaines usuelles. Les résultats numériques obtenus montrent l'extraordinaire maîtrise de Mayer.*

MOTS CLÉS – Algorithme des transferts de Régnier, Classification, Histoire de la statistique, Méthode des moyennes de Mayer, Régression linéaire

SUMMARY – Mayer's fitting method and its links to clustering methods
The simple case of Mayer's straight line fitting, which was taught in French secondary schools some years ago, was introduced in some curricula as a surrogate to least-squares. It turns out that the procedure thus proposed to secondary school students provides a basic example of a regression tree. It also turns out, in the general case, that it is a clustering problem for which Régnier's transfer algorithm [1965] is well suited, albeit possibly suboptimal. The famous example of fitting which Mayer treated in 1750 by an innovative and general method is revisited in the light of standard present-day statistical methods. The numerical results show the outstanding expertise of Mayer.

KEYWORDS – Clustering, History of statistics, Linear regression, Mayer's method of averages, Régnier's transfer algorithm

1. INTRODUCTION

Cherchant à résoudre un système d'équations linéaires numériquement incompatibles, l'astronome Tobias Mayer (1723–1762) propose de sommer (ou moyenner) ces équations par groupe en définissant autant de groupes disjoints d'observations qu'il y a de coefficients à estimer. Il résout alors, lorsque c'est le cas, le système de Cramer[2] ainsi défini. La méthode, publiée par Mayer (cf. [Mayer, 1750] exige donc

[1]Laboratoire de statistique et probabilités, Institut de mathématiques, Université Paul Sabatier (Toulouse III), 31062 Toulouse cedex 9, antoine.falguerolles@math.univ-toulouse.fr

[2]Gabriel Cramer (1704-1752), mathématicien suisse, est un contemporain de l'astronome allemand Tobias Mayer.

qu'une partition soit fournie *a priori* mais son auteur ne propose pas de procédure générale permettant de guider le choix de cet élément décisif.

L'analyse faite par Mayer de données d'observation est classiquement présentée comme un prototype de la régression que nous connaissons actuellement : les estimations sont obtenues en résolvant un système d'équations linéaires qui rappelle celui formé par les équations normales de la méthode des moindres carrés [Legendre, 1805]. Dans cet article, la méthode de Mayer est considérée comme un problème de régression sous contrainte de classification. Différents critères d'optimisation pour la recherche de partitions sont introduits et, parmi ceux-ci, l'inusable critère des moindres carrés. Un algorithme de transfert, analogue à celui introduit par Simon Régnier (1932-1980) pour rechercher une partition centrale d'une famille de partitions [Régnier, 1965], permet d'affiner des solutions initiales obtenues par des méthodes usuelles de la classification automatique.

Cet article est organisé comme suit. La méthode d'ajustement de Mayer est d'abord rappelée : construction de la droite de Mayer pour la régression linéaire simple (Section 2), construction générale de Mayer pour la régression linéaire multiple (Section 3). C'est l'ordre pédagogique convenu qui est couramment adopté dans l'enseignement de la régression bien que Mayer, comme d'ailleurs Legendre pour la méthode des moindres carrés [Legendre, 1805], ait considéré d'emblée le cas multiple. Les principales propriétés des estimateurs de Mayer sont alors étudiées. Puis, les aspects classificatoires sont présentés en Section 4 : sans surprise, l'algorithme de transfert de Régnier trouve là une nouvelle sphère d'application. Enfin, l'exemple le plus connu de Mayer est reconsidéré (Section 5).

2. CONSTRUCTION DE LA DROITE DE RÉGRESSION SELON MAYER

Rappelons que cette méthode avait été introduite dans l'enseignement secondaire en remplacement de celle dite des « moindres carrés », vraisemblablement abandonnée en raison de sa prétendue complexité[3]. De fait, la méthode de Mayer, assez intuitive pour l'ajustement d'une régression linéaire simple, présente un certain nombre de propriétés qui justifient pleinement son introduction dans certains programmes officiels d'enseignement.

2.1. CONSTRUCTION

Soient X et Y deux variables statistiques quantitatives conjointement observées. Soit $\{(x_i, y_i)|1 \leq i \leq n\}$ le nuage associé des observations. Le problème de la régression linéaire de la variable réponse Y sur une variable explicative X, avec constante, est celui de l'estimation des coefficients de l'espérance conditionnelle

$$\mu(x) = E[Y|X = x] = \beta_0 + \beta_1 x.$$

[3]Il faut bien reconnaître que les notations couramment utilisées ajoutent à la confusion des lycéens : coefficients inconnus notés a, b, ... et valeurs observées connues x, y, ... alors que, dans sa publication de la méthode des moindres carrés Legendre [1805] notait les coefficients inconnus x, y, ... et les valeurs observées connues a, b, ... Le calcul de sommes de doubles-produits et de carrés que la méthode exige est aussi une source potentielle d'erreurs numériques.

C'est l'équation d'une droite dont il faut estimer les coefficients β_0 et β_1. L'idée de Mayer est de construire un système de deux équations à deux inconnues dont la solution unique fournit les estimations cherchées. Ceci revient à déterminer les coordonnées de deux points distincts résumant les observations puis à calculer l'ordonnée à l'origine et la pente de la droite passant par ces deux points.

La méthode de Mayer, telle qu'enseignée en France[4], consiste à se donner un seuil s_X, $\min\{x_i\} < s_X < \max\{x_i\}$, et à construire la droite passant par les deux points de coordonnées

$$(\overline{x[x < s_X]}, \overline{y[x < s_X]}) \text{ et } (\overline{x[x \geq s_X]}, \overline{y[x \geq s_X]}),$$

$\overline{\bullet[condition]}$ désignant la moyenne conditionnelle empirique d'une variable notée $\bullet$. Classiquement, $\overline{x}$ et $\overline{y}$ désignent les moyennes empiriques usuelles.

Il est alors facile de vérifier que :

1. le point moyen $(\overline{x}, \overline{y})$ du nuage appartient à la droite de Mayer ;
2. les estimations b_0 et b_1 des coefficients β_0 et β_1 sont données par des formules simples,

$$\left\{ \begin{array}{rcl} b_0 & = & \dfrac{\overline{x[x \geq s_X]}\ \overline{y[x < s_X]} - \overline{x[x < s_X]}\ \overline{y[x \geq s_X]}}{\overline{x[x \geq s_X]} - \overline{x[x < s_X]}} \\ b_1 & = & \dfrac{\overline{y[x \geq s_X]} - \overline{y[x < s_X]}}{\overline{x[x \geq s_X]} - \overline{x[x < s_X]}} \ . \end{array} \right.$$

L'estimation m_i de $\mu(x_i)$, la moyenne ajustée de l'observation i, vaut alors

$$m_i = b_0 + b_1 x_i = \overline{y} + b_1(x_i - \overline{x}) \ .$$

Il est clair que le choix, tout aussi arbitraire, des points de coordonnées

$$(\overline{x[x \leq s_X]}, \overline{y[x \leq s_X]}) \text{ et } (\overline{x[x > s_X]}, \overline{y[x > s_X]}),$$

plutôt que des points

$$(\overline{x[x < s_X]}, \overline{y[x < s_X]}) \text{ et } (\overline{x[x \geq s_X]}, \overline{y[x \geq s_X]}),$$

fournit encore des estimations voisines ayant des propriétés identiques[5].

Dans la pratique, le seuil s_X est posé égal à $\overline{x}$. Mais tout autre indice de position des valeurs x, une médiane par exemple, serait aussi acceptable comme on le verra dans la sous-section 2.2.

Enfin, on notera que, comme pour les moindres carrés, les résultats ne présentent pas une propriété d'invariance par rapport au choix de la variable réponse.

[4]Il semble que, dans la seconde moitié du XX^{e} siècle, la France était le seul pays où la méthode de Mayer ait figuré dans des programmes officiels d'enseignement.

[5]Il existe encore une variante de la méthode de Mayer dans laquelle :

$$\left\{ \begin{array}{rcl} b_1 & = & \dfrac{\overline{y[x > s_X]} - \overline{y[x < s_X]}}{\overline{x[x > s_X]} - \overline{x[x < s_X]}} \\ b_0 & = & \overline{y} - b_1 \overline{x} \ . \end{array} \right.$$

Ceci revient à éliminer dans les observations pour lesquelles $x_i = \overline{x}$. Le choix arbitraire du sens de l'inégalité à prendre au sens strict ne se pose alors plus.

2.2. À l'ombre de la droite de Mayer : un arbre de régression !

Le principe de construction de la droite de Mayer n'est pas sans évoquer celui de la construction d'un arbre de régression[6] à un seul niveau :

$$\begin{array}{lll} \textbf{si} \quad x_i < s_X & \textbf{alors} & m_i = \overline{y[x < s_X]} \\ & \textbf{sinon} & m_i = \overline{y[x \geq s_X]}. \end{array}$$

On sait que, dans cette approche, c'est la maximisation de la variance expliquée qui détermine le choix de la valeur du seuil s_X. Cette variance vaut ici :

$$\frac{\sum_{i=1}^n \mathbf{1}(x_i < s_X) \sum_{i=1}^n \mathbf{1}(x_i \geq s_X)}{n^2} \; (\; \overline{y[x \geq s_X]} \; - \; \overline{y[x < s_X]} \;)^2$$

où $\mathbf{1}(\bullet)$ désigne l'indicatrice de la condition $\bullet$.

Cette expression permet de vérifier qu'il n'existe pas de solution générale pour le choix du seuil s_X. Une partie de la formule, soit

$$\frac{\sum_{i=1}^n \mathbf{1}(x_i < s_X) \sum_{i=1}^n \mathbf{1}(x_i \geq s_X)}{n^2},$$

suggère le choix de la médiane : cette quantité est maximum lorsque

$$\sum_{i=1}^n \mathbf{1}(x_i < s_X) = \frac{n}{2}$$

(en négligeant l'aspect discret du problème). L'autre partie, soit

$$(\; \overline{y[x \geq s_X]} \; - \; \overline{y[x < s_X]} \;)^2,$$

n'indique rien de très précis ; sa valeur dépend surtout de la liaison entre X et Y : cette quantité est d'autant plus grande que, en moyenne, les valeurs de la réponse y correspondant à des valeurs de la variable explicative x supérieures au seuil s_X sont plus grandes (respectivement plus petites) que celles correspondant à des valeurs strictement inférieures au seuil s_X.

3. RÉGRESSION LINÉAIRE MULTIPLE ET MÉTHODE DE MAYER

Comment se généralise l'approche de Mayer au cas de la régression linéaire multiple ? Quelles sont les propriétés statistiques des estimateurs correspondants ? Revenons d'abord sur l'exemple historique de Mayer.

[6] Un arbre de régression est un arbre binaire (au sens de la théorie des graphes). Il représente les conditions sur les variables explicatives conduisant à une partition de l'échantillon en classes d'unités statistiques. Ces classes sont déterminées en sorte que la variable réponse, quantitative, présente des moyennes intra-classes différenciées et des variances intra-classes les plus petites possibles.

3.1. L'EXEMPLE DE TOBIAS MAYER

L'exemple qui a fait connaître la méthode de Mayer a trait à la mesure de la libration lunaire. Les données consistent en 27 observations du cratère lunaire Manilius effectuées pendant une année. Partant d'une relation théorique liant les mesures de trois arcs à trois coefficients inconnus par des formules de trigonométrie sphérique, Mayer établit, par linéarisation[7], la relation suivante :

$$\beta - (90^\circ - h) \quad = \quad \alpha \sin(g-k) - \alpha \sin(\theta)\cos(g-k)$$

où g et h sont des données observées, k est obtenu dans des tables établies par Euler[8], et β, α et $\sin(\theta)$ les coefficients inconnus (cf. [Farebrother, 1998, chapitre 1, p. 11-15 ; Hald, 2007, chapitre 6, p. 48-49]). Les données, reproduites dans l'ouvrage de Farebrother [1998, p. 14], sont rappelées dans le Tableau 1 ci-après.

Le problème étudié par Mayer consistait donc à estimer les paramètres inconnus β, α et $\sin\theta$ à partir de 27 observations, donc à partir d'un système de 27 équations linéaires à 3 coefficients inconnus (β, α et $\alpha \sin\theta$). Si la relation avait été exacte, tout choix d'un sous-système de rang 3 aurait donné les mêmes valeurs des inconnues. Mais compte tenu des erreurs d'observations et des approximations du modèle, force est de constater que :

$$\beta - (90^\circ - h) \quad \approx \quad \alpha \sin(g-k) - \alpha \sin(\theta)\cos(g-k).$$

Le système constitué par les 27 équations linéaires est donc incompatible, et Mayer est confronté au problème ancien de recherche d'une solution de compromis.

Toutefois, le problème traité par Mayer peut être formulé comme celui d'une régression multiple dans laquelle la variable réponse observée serait $y = (90^\circ - h)$, les deux variables explicatives $x_1 = -\sin(g-k)$ et $x_2 = \cos(g-k)$, et les coefficients

[7]La démarche générale peut être formalisée comme suit. Soit une fonction $F_{\underline{\alpha}}(\underline{z}) = 0$ spécifiant une relation théorique liant un vecteur réel $\underline{\alpha}$ de coefficients inconnus de dimension p et un vecteur réel $\underline{z}$ de grandeurs observables de dimension q. Le savoir-faire des grands précurseurs réside alors dans une linéarisation *ad hoc* du problème en vue de l'estimation de $\underline{\alpha}$. Schématiquement, la méthode consiste en la détermination

- d'une application bijective ϕ de $\mathbb{R}^p$ dans $\mathbb{R}^p$ permettant de reparamétriser le problème ($\underline{\beta} = \phi(\underline{\alpha})$, $\underline{\alpha} = \phi^{-1}(\underline{\beta})$),
- de $p+1$ applications ψ_j, $j \in \{0, \ldots, p\}$, de $\mathbb{R}^q$ dans $\mathbb{R}$, transformant les données d'observation primaires en données d'observation secondaires,

telles que

$$\underline{\beta} : \ \psi_0(\underline{z}) \approx \sum_{j=1}^{p} \beta_j \psi_j(\underline{z}) \quad \Rightarrow \quad F_{\underline{\alpha}}(\underline{z}) \approx 0 \ \text{ avec } \ \alpha = \phi^{-1}(\underline{\beta}).$$

Reste alors à estimer les coefficients $\underline{\beta}$ en combinant n ($n \geq p$) observations de $\underline{z}$.

C'est la démarche de Mayer dans son étude la libration de la Lune en 1750 (cf. [Mayer, 1750]) ; mais c'est aussi de nombreux autres grands auteurs pour des situations très variées (cf. [Hald, 2007, chapitre 6]) : Boscovich, Laplace, Legendre ... Bien sûr, il resterait à préciser le sens mathématique de l'approximation dans les formules ci-dessus. De plus, dans cette approche, le choix de la variable $\psi_0(\underline{z})$ en tant que variable réponse dans une problématique de type régression reste quelque peu artificiel.

[8]Leonhard Euler (1707–1783), mathématicien suisse, est un contemporain de Tobias Mayer.

TABLEAU 1. Tableau des données de Mayer reproduites d'après l'ouvrage de Farebrother [1998, p. 14]. La première colonne donne le numéro de l'observation et la dernière l'étiquette de la classe considérée par Mayer dans son problème d'estimation. Les mesures des angles sont exprimées en degrés.

	$(90° - h)$	$\sin(g-k)$	$\cos(g-k)$	classe
1	-13° 10	0,8836	-0,4682	1
2	-13° 8	0,9996	-0,0282	1
3	-13° 12	0,9899	0,1421	1
4	-14° 15	0,2221	0,9750	3
5	-14° 42	0,0006	1,0000	3
6	-13° 1	0,9308	-0,3654	1
7	-14° 31	0,0602	0,9982	3
8	-14° 57	-0,1570	0,9876	2
9	-13° 5	0,9097	-0,4152	1
10	-13° 2	1,0000	0,0055	1
11	-13° 12	0,9689	0,2476	1
12	-13° 11	0,8878	0,4602	1
13	-13° 34	0,7549	0,6558	3
14	-13° 53	0,5755	0,8178	3
15	-13° 58	0,3608	0,9326	3
16	-14° 14	0,1302	0,9915	3
17	-14° 56	-0,1068	0,9943	3
18	-14° 47	-0,3363	0,9418	2
19	-15° 56	-0,8560	0,5170	2
20	-13° 29	0,8002	0,5997	3
21	-15° 55	-0,9952	-0,0982	2
22	-15° 39	-0,8409	0,5412	2
23	-16° 9	-0,9429	0,3330	2
24	-16° 22	-0,9768	0,2141	2
25	-15° 38	-0,6262	-0,7797	2
26	-14° 54	-0,4091	-0,9125	2
27	-13° 7	0,9284	-0,3716	1

de la régression $\beta_0 = \beta$, $\beta_1 = -\alpha$, et $\beta_2 = \alpha \sin(\theta)$. Le modèle de régression est alors :

$$E[Y|X_1 = x_1, X_2 = x_2] \quad = \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Le principe de la solution proposée par Mayer consiste à considérer une partition des observations en 3 classes (le nombre d'inconnues) de mêmes effectifs ; puis, à sommer (ou moyenner) les observations (les équations) au sein de chaque classe et résoudre le système de Cramer ainsi obtenu.

Comment construire cette partition ? Hald [2007, p. 49] rapporte que Mayer, à bon escient, indique que les moyennes par classes doivent être le plus différentes possibles. Pour ce faire, Mayer exploite la structure particulière de son problème. D'une part, les observations de la variable x_1 sont plus dispersées que celles de la variable x_2. (De façon tout à fait anachronique, on peut vérifier que l'écart type empirique de x_1 est supérieur à celui de x_2.) D'autre part les variables x_1 et x_2 sont liées par la relation $x_1^2 + x_2^2 = 1$. De ce point de vue, le problème est unidimensionnel. De fait,

deux terciles (quantiles d'ordres respectifs $\frac{1}{3}$ et $\frac{2}{3}$) de la distribution empirique de x_1 définissent la partition en trois classes proposée par Mayer. Mais, à l'évidence, la solution *ad hoc* utilisée dans cet exemple n'a pas de portée générale.

De plus, on voit bien que la méthode appliquée dans le cas simple (une seule variable explicative et un seuil de coupure) ne se généralise pas directement au cas multiple (ici deux variables explicatives). La partition qui serait obtenue en croisant les bipartitions engendrées pour chaque variable explicative (ici 2 bipartitions) risque de ne pas comporter un nombre de classes non vides (ici au plus 2^2) exactement égal à celui du nombre de coefficients à estimer (ici 3).

3.2. ARTICULATION ENTRE RÉGRESSION ET PARTITION

Soit donc le problème général de régression dans lequel on considère un vecteur réponse aléatoire $\underline{Y}$ de $\mathbb{R}^n$, de matrice expérimentale $\mathbf{X}$ (de dimension (n, p), $n \geq p$, et supposée de rang p). Les caractéristiques distributionnelles supposées sont :

- vecteur moyenne : $E[\underline{Y}] = \mathbf{X}\underline{\beta}$ où $\underline{\beta}$ est le vecteur des p coefficients inconnus de la régression,
- matrice de variance : $Var(\underline{Y}) = \sigma^2\mathbf{I_n}$ où $\mathbf{I_n}$ désigne la matrice identité et σ^2 un réel positif connu ou inconnu.

On désigne par $\underline{y}$ la réalisation considérée de la variable aléatoire $\underline{Y}$, et par $\underline{b}$ (resp. $\underline{m}$) l'estimation du vecteur des coefficients inconnus $\underline{\beta}$ (resp. du vecteur des moyennes inconnues $\underline{\mu}$) fournie par un estimateur $\widehat{\underline{\beta}}$ (resp. $\widehat{\underline{\mu}}$). Il va de soi que l'on s'intéresse ici principalement à l'estimateur de Mayer.

Soit alors une partition c de l'ensemble $\{1, \ldots, n\}$ en exactement p classes, $\mathbf{C}$ la matrice de dimension (n, p) des indicatrices des classes de cette partition et $\mathbf{C}'$ sa transposée. De façon générale, $\mathbf{U}'$ (resp. $\underline{u}'$) désignera la transposée d'une matrice $\mathbf{U}$ (resp. d'un vecteur $\underline{u}$). La méthode générale de Mayer consiste à agréger les données au sein de chaque classe et donc à résoudre le système :

$$\mathbf{C}'\mathbf{X}\underline{b} \;=\; \mathbf{C}'\underline{y}\ .$$

La régularité de la matrice $\mathbf{C}'\mathbf{X}$ garantit l'unicité de la solution : $\underline{b} = (\mathbf{C}'\mathbf{X})^{-1}\mathbf{C}'\underline{y}$. Sous cette réserve, les estimateurs de Mayer $\widehat{\underline{\beta}}$ et $\widehat{\underline{\mu}}$ de $\underline{\beta}$ et $\underline{\mu}$ sont donc donnés par des formules simples :

$$\begin{aligned}\widehat{\underline{\beta}} &= (\mathbf{C}'\mathbf{X})^{-1}\mathbf{C}'\underline{Y}\\ \widehat{\underline{\mu}} &= \mathbf{X}\widehat{\underline{\beta}}.\end{aligned}$$

Ces estimateurs, linéaires par construction, possèdent la propriété classique d'absence de biais.

Enfin, leurs matrices de variance ont également des expressions simples. Si l'on désigne par $n_1, \ldots, n_p$ ($n = n_1 + \ldots + n_p$) les effectifs des p classes de la partition considérée et par $\mathbf{diag}(n_1, \ldots, n_p)$ la matrice diagonale ayant pour éléments ces effectifs $n_1, \ldots, n_p$, il vient $\mathbf{C}'\mathbf{C} = \mathbf{diag}(n_1, \ldots, n_p)$. On obtient alors :

$$\begin{aligned}\mathrm{Var}(\widehat{\underline{\beta}}) &= \sigma^2(\mathbf{C}'\mathbf{X})^{-1}\ \mathbf{diag}(n_1, \ldots, n_p)\ (\mathbf{X}'\mathbf{C})^{-1}\\ \mathrm{Var}(\widehat{\underline{\mu}}) &= \mathbf{X}\ \mathrm{Var}(\widehat{\underline{\beta}})\ \mathbf{X}'.\end{aligned}$$

Notons que si la matrice de variance du vecteur aléatoire réponse n'était pas diagonale mais de la forme plus générale $\sigma^2\mathbf{\Sigma}$, alors :

$$\mathrm{Var}(\widehat{\underline{\beta}}) \;=\; \sigma^2(\mathbf{C}'\mathbf{X})^{-1}\,\mathbf{C}'\mathbf{\Sigma}\mathbf{C}\,(\mathbf{X}'\mathbf{C})^{-1}.$$

3.3. OPTIMALITÉ

Les estimateurs de Mayer ne sont pas sans rappeler ceux, plus connus, fournis par la méthode des moindres carrés :

$$\begin{aligned}\widehat{\underline{\beta^\star}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{Y} \qquad (\ Var(\widehat{\underline{\beta^\star}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\ )\\ \widehat{\underline{\mu}^\star} &= \mathbf{X}\widehat{\underline{\beta^\star}} \qquad (\ Var(\widehat{\underline{\mu}^\star}) = \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ ).\end{aligned}$$

Les estimateurs des moindres carrés (Legendre) partagent avec ceux de Mayer les propriétés de linéarité et d'absence de biais. Mais, les premiers (Legendre) présentent des propriétés d'optimalité de la variance établies par le théorème de Gauss-Markov (cf. par exemple, [Tassi, 1989, p. 358-359]) que ne présentent pas les derniers (Mayer).

Mieux encore, en retenant le critère des moindres carrés, on peut encore quantifier le caractère éventuellement sous-optimal d'une solution de Mayer. En effet, notant respectivement $\|\bullet\|_2$ et $\|\bullet\|_{\mathbf{M}}$ la norme euclidienne usuelle et la norme associée à une matrice définie positive $\mathbf{M}$ d'un vecteur réel $\bullet$ de $\mathbb{R}^q$, il vient :

$$\begin{aligned}\|\underline{Y}-\widehat{\mu}\|_2^2 &= \|\underline{Y}-\widehat{\underline{\mu}^\star}\|_2^2 + \|\widehat{\underline{\mu}^\star}-\widehat{\underline{\mu}}\|_2^2\\ &= \|\underline{Y}-\widehat{\underline{\mu}^\star}\|_2^2 + \|\mathbf{X}\widehat{\underline{\beta^\star}}-\mathbf{X}\widehat{\underline{\beta}}\|_2^2\\ &= \|\underline{Y}-\widehat{\underline{\mu}^\star}\|_2^2 + \|\widehat{\underline{\beta^\star}}-\widehat{\underline{\beta}}\|_{\mathbf{X}'\mathbf{X}}^2\end{aligned}$$

En terme d'ajustement, le prix à payer pour une construction des estimateurs faisant appel à une partition est alors $0 \leq \|\widehat{\underline{\mu}^\star}-\widehat{\underline{\mu}}\|_2^2 = \|\widehat{\underline{\beta^\star}}-\widehat{\underline{\beta}}\|_{\mathbf{X}'\mathbf{X}}^2$. Et rien ne permet d'affirmer qu'il existe nécessairement une partition pour laquelle estimateurs de Mayer et estimateurs des moindres carrés coïncident.

D'ailleurs, pour évaluer la pénalité associée à une partition admissible, ce seront des quantités analogues à celle de l'estimation $\frac{\|\underline{y}-\underline{m}\ \|_2}{\|\underline{y}-\underline{m}^\star\|_2}$ (exprimée en %) du rapport $\frac{\|\underline{Y}-\widehat{\underline{\mu}}\|_2}{\|\underline{Y}-\widehat{\underline{\mu}^\star}\|_2}$ qui seront retenues dans la section 5.

4. UN PROBLÈME DE CLASSIFICATION

La méthode générale de Mayer appelle au moins deux remarques. D'abord, les estimations fournies dépendent du choix de la partition en p classes non vides qui était considérée comme donnée dans la section précédente. Comment donc choisir une bonne partition ? Ensuite, à l'opposé d'un certain nombre de méthodes usuelles de régression (moindres carrés, moindres valeurs absolues, moindre médiane des carrés, moindres carrés élagués ...[9]), la méthode de Mayer ne se réfère pas explicitement à

[9]Curieusement, il n'existe pas d'ouvrage recensant toutes les méthodes de régression. En l'état actuel, les sources d'information les plus facilement accessibles sont les livres d'histoire de la statistique (cf. par exemple [Stigler, 1986 ; Farebrother, 1998 ; Stigler, 1999 ; Hald, 2007]) et les manuels associés à d'utilisation des logiciels statistiques (cf. par exemple [Venables et Ripley, 2002]).

un critère global d'optimisation. De fait, l'intérêt peut être porté plus sur l'estimation $\underline{b}$ des coefficients inconnus $\underline{\beta}$ de la régression que sur l'estimation $\underline{m} = \mathbf{X}\underline{b}$ des moyennes $\underline{\mu} = \mathbf{X}\underline{\beta}$. Ou inversement. On sent donc bien que le choix d'une partition sur laquelle s'appuie la méthode de Mayer correspond au choix implicite d'un critère d'optimisation qu'il faut chercher à expliciter. Dans ce qui suit, on note par $\mathcal{C}_{\mathbf{X},p}$ l'ensemble des partitions c de n observations en p classes permettant d'appliquer la méthode de Mayer ; autrement dit, de partitions c telles que les matrices $\mathbf{C}$ des indicatrices de leurs classes vérifient la propriété de régularité du produit matriciel $\mathbf{C}'\mathbf{X}$, $\mathbf{C}'$ désignant la transposée de la matrice $\mathbf{C}$. Dans ce qui suit, ces partitions sont dites admissibles.

4.1. CRITÈRES D'OPTIMISATION

Divers critères d'optimisation peuvent être retenus. On peut les ranger en trois catégories selon l'objectif statistique poursuivi.

4.1.1. Catégorie 1

On s'intéresse essentiellement à la qualité de l'ajustement global du modèle de régression. Il faut alors quantifier l'écart entre valeur observée de la réponse, $\underline{y}$, du vecteur aléatoire réponse et estimation de sa moyenne, $\underline{m}$. Pour marquer que cette estimation est intimement liée au choix d'une partition admissible c on la note $\underline{m}(c)$.

Soit alors une application F de $\mathbb{R}^n \times \mathbb{R}^n$ dans $[0, +\infty[$ exprimant la qualité globale de cet ajustement. Si l'on désigne par $\| \bullet \|_2$ et $\| \bullet \|_1$ respectivement, les normes L_2 et L_1 d'un vecteur $\bullet$ de $\mathbb{R}^n$, des choix usuels sont $F(\underline{y}, \underline{m}(c)) = \|\underline{y} - \underline{m}(c)\|_2^2$, ou $F(\underline{y}, \underline{m}(c)) = \|\underline{y} - \underline{m}(c)\|_1$. Mais, il existe bien d'autres choix moins classiques : la médiane des carrés des coordonnées du vecteur $\underline{y} - \underline{m}(c)$, le maximum des carrés des coordonnées du vecteur $\underline{y} - \underline{m}(c)$ ...

Le problème consiste alors à rechercher une partition $c^\star$ appartenant à l'ensemble $\mathcal{C}_{\mathbf{X},p}$ telle que :

$$c^\star \in \arg \min_{c \in \mathcal{C}_{\mathbf{X},p}} F(\underline{y}, \underline{m}(c))$$

4.1.2. Catégorie 2

Dans l'exemple traité par Mayer, l'accent est mis plus sur l'estimation des coefficients du modèle que sur l'estimation des moyennes ajustées. D'ailleurs, dans certaines situations, la valeur théorique de ces coefficients est déjà connue par d'autres moyens[10]. On s'intéresse alors essentiellement à l'écart entre l'estimation obtenue $\underline{b}(c)$ associée à une partition admissible et une valeur préspécifiée $\underline{b}_0$ du vecteur des coefficients.

Soit maintenant une application H de $\mathbb{R}^p \times \mathbb{R}^p$ dans $[0, +\infty[$ exprimant la proximité de ces vecteurs de coefficients. Ces derniers étant souvent exprimés dans dans des unités très dissemblables, les métriques usuelles, $\|\underline{b}_0 - \underline{b}(c)\|_2^2$ ou $\|\underline{b}_0 - \underline{b}(c)\|_1$,

[10] C'est, en particulier, le cas dans l'exemple publié en 1805 par Legendre pour illustrer la méthode des moindres carrés [Falguerolles et Pinchon, 2006].

peuvent s'avérer inadaptées. Il semble, à cet égard, préférable de retenir une métrique de type Mahalanobis[11].

On recherche alors une partition $c^\star$ de l'ensemble $\mathcal{C}_{\mathbf{X},p}$ des partitions c en p classes non vides telle que :

$$c^\star \in \arg\ \min_{c\in\mathcal{C}_{\mathbf{X},p}} H(\underline{b}_0, \underline{b}(c))$$

4.1.3. Catégorie 3

On peut aussi vouloir combiner les deux critères F et H. Dans cet esprit, la démarche dite *lasso* [Tibshirani, 1996]) semble tout à fait indiquée. Les variables explicatives étant préalablement centrées et réduites, on recherche une partition $c^\star$ de l'ensemble $\mathcal{C}_{\mathbf{X},p}$ telle que :

$$\begin{aligned} c^\star \in \ & \arg\ \min_{c\in\mathcal{C}_{\mathbf{X},p}} \|\underline{y} - \underline{m}(c)\|_2^2 \\ & \text{sous la contrainte} \\ & \|\underline{b}_0 - \underline{b}(c)\|_1 \leq t \end{aligned}$$

où t est un paramètre de réglage de la méthode.

4.2. Une réponse localement optimale

Les problèmes d'optimisation définis ci-dessus ne sont pas usuels en statistique. Leur particularité, qui en fait d'ailleurs leur spécificité, provient de la nature particulière des contraintes : $c \in \mathcal{C}_{\mathbf{X},p}$. Mais, bien que cet ensemble soit de cardinal fini, il reste hors de question, vu sa taille, de l'explorer systématiquement.

L'emploi de l'algorithme des transferts de Régnier semble ici tout à fait indiqué[12]. En effet, supposons qu'à l'étape courante on dispose d'une partition admissible c. On recherche alors s'il existe un transfert d'un élément d'une classe dans une autre qui fournisse une partition admissible améliorant la valeur du critère à optimiser. Le cas échéant, on effectue le transfert considéré définissant une nouvelle partition admissible c' ; on réitère alors la recherche d'un nouveau transfert améliorant. Sinon, la partition c est déclarée localement optimale.

On conçoit que l'algorithme soit assez lent puisque l'on doit considérer, à chaque étape, jusqu'à $p-1$ transferts pour chacun des n éléments. Toutefois les calculs à effectuer restent assez simples comme montré ci-dessous.

[11]Ces métriques sont associées à l'inverse d'une matrice de variance. Dans la sous-section 3.3, la matrice de variance de $\widehat{\underline{\beta}^\star}$ était $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, et la quantité $\|\widehat{\underline{\beta}^\star} - \widehat{\underline{\beta}}\|^2_{\mathbf{X}'\mathbf{X}}$ était naturellement apparue comme forme équivalente de $\|\widehat{\underline{\mu}^\star} - \widehat{\underline{\mu}}\|_2^2$. Dans le même esprit, on peut considérer $\|\widehat{\underline{\beta}^\star} - \widehat{\underline{\beta}}\|^2_{(diag((\mathbf{X}'\mathbf{X})^{-1}))^{-1}}$, $diag(\bullet)$ désignant la matrice diagonale constituée à partir des éléments diagonaux de la matrice carrée $\bullet$. Dans ces deux exemples, les coordonnées des vecteurs $\underline{\beta}$ sont, de fait, standardisées et ne dépendent donc plus des unités dans lesquelles elles s'expriment.

[12]Rappelons que dans son article de 1965, Simon Régnier résout un des premiers problèmes d'analyse de données symboliques [Bock et Diday, 2000] : il s'attaque au problème de la détermination d'une partition moyenne (centrale) d'une série statistique de partitions, comme un statisticien classique recherche le milieu (une moyenne) d'une série statistique quantitative classique.

Soit une partition admissible courante c de matrice indicatrice $\mathbf{C}$ et les estimations $\underline{b}(c)$ (coefficients de la régression) et $\underline{m}(c)$ (moyennes ajustées) associées à cette partition par la méthode de Mayer. Désignons par $\mathbf{G_c}$ la matrice $\mathbf{C}'\mathbf{X}$ et par $\underline{g}_c$ le vecteur $\mathbf{C}'\underline{y}$ en sorte que $\underline{b}(c) = {\mathbf{G_c}}^{-1}\underline{g}_c$ (et $\underline{m}(c) = \mathbf{X}\ \underline{b}(c)$). Sans perte de généralité, considérons alors le transfert de l'élément i, $i \in \{1, \ldots, n\}$, de la classe étiquetée 1 vers la classe étiquetée 2. Soit $c_\bullet$ la partition ainsi construite.

Soient $\underline{\mathbf{1}}_i$ l'indicatrice de l'élément i (le vecteur de dimension n tel que $\underline{\mathbf{1}}_i' = (0, \ldots, 0, 1, 0, \ldots, 0)$), $\underline{x}_i$ le i^{e} vecteur-ligne de la matrice $\mathbf{X}$ ($\underline{x}_i = \mathbf{X}'\underline{\mathbf{1}}_i$) et $\underline{t}_{\ 1\to 2}$ le vecteur de dimension p associé au transfert de l'élément i de la classe 1 dans la classe 2 ($\underline{t}'_{\ 1\to 2} = (-1, 1, 0, \ldots, 0)$).

Du calcul matriciel classique[13] permet de vérifier que ce transfert, admissible si et seulement si $1 + \underline{x}_i'{\mathbf{G_c}}^{-1}\underline{t}_{\ 1\to 2} \neq 0$, aboutit à réviser simplement les éléments de calcul fournissant les estimations courantes :

$$\begin{aligned}
\underline{b}(c_\bullet) &= \underline{b}(c) + \frac{1}{1 + \underline{x}_i'{\mathbf{G_c}}^{-1}\underline{t}_{\ 1\to 2}}\ (y_i - m(c)_i){\mathbf{G_c}}^{-1}\underline{t}_{\ 1\to 2} \\
\underline{m}(c_\bullet) &= \underline{m}(c) + \frac{1}{1 + \underline{x}_i'{\mathbf{G_c}}^{-1}\underline{t}_{\ 1\to 2}}\ (y_i - m(c)_i)\mathbf{X}{\mathbf{G_c}}^{-1}\underline{t}_{\ 1\to 2}
\end{aligned}$$

Si l'on choisit un critère d'optimisation dans l'une des catégories définies dans la sous-section 4.1, il est donc numériquement aisé de rechercher pas à pas un transfert améliorant la valeur du critère d'optimisation retenu et, le cas échéant, de modifier la partition couramment considérée en effectuant le transfert identifié.

Enfin, toutes les formules (tant celles de la définition générale que celles des transferts élémentaires) montrent que la méthode se prête aussi à l'introduction de poids pour les observations. Mais cette flexibilité n'est pas utilisée dans cet article.

4.3. RECHERCHE D'UNE PARTITION INITIALE

Il s'agit ici d'initialiser la procédure de transferts en proposant une partition admissible permettant d'amorcer la procédure itérative décrite à la sous-section 4.2. À cette fin, trois catégories de méthodes semblent particulièrement adaptées :

- Des méthodes *ad hoc* comme celle utilisée par Mayer (cf. sous-section 3.1).
- Des méthodes de type centres mobiles encore appelées méthodes des K-moyennes, traduction littérale de l'anglo-saxon *K-means*, où la lettre K rappelle que la méthode a pour objet la construction d'une partition d'un ensemble en K classes. Ici $K = p$.
- Des méthodes de type arbre de régression, le nombre de « feuilles » étant alors fixé a priori et posé égal au nombre p de coefficients à estimer.

[13] On a $\mathbf{C_{c_\bullet}} = \mathbf{C_c} + \underline{\mathbf{1}}_i\ \underline{t}'_{\ 1\to 2}$ d'où $\mathbf{G_{c_\bullet}} = \mathbf{G_c} + \underline{t}_{\ 1\to 2}\ \underline{x}_i'$ et $g_{c_\bullet} = g_c + y_i\ \underline{t}_{\ 1\to 2}$. Il suffit alors d'appliquer la formule donnant l'inverse de la matrice $\mathbf{G_c} + \underline{t}_{\ 1\to 2}\ \underline{x}_i'$ en fonction de l'inverse ${\mathbf{G_c}}^{-1}$ de la matrice $\mathbf{G_c}$, et des vecteurs $\underline{t}_{\ 1\to 2}$ et $\underline{x}_i$:

$$(\mathbf{G_c} + \underline{t}_{\ 1\to 2}\ \underline{x}_i')^{-1} = {\mathbf{G_c}}^{-1} - \frac{1}{1 + \underline{x}_i'{\mathbf{G_c}}^{-1}\underline{t}_{\ 1\to 2}}{\mathbf{G_c}}^{-1}\underline{t}_{\ 1\to 2}\ \underline{x}_i'{\mathbf{G_c}}^{-1}.$$

5. RETOUR SUR L'EXEMPLE DE MAYER

On cherche ici à illustrer l'application de certaines des méthodes présentées ci-dessus sur l'exemple traité par Mayer et à étudier leur bien-fondé. On va donc chercher à évaluer ainsi les pénalités encourues en utilisant une estimation basée sur une partition admissible plutôt que sur une méthode de régression sans contrainte.

La régression sans contrainte, comme d'ailleurs la méthode des transferts, exige un critère d'optimisation. Trois critères, des plus connus et appartenant à la catégorie 1 étudiée dans la sous-section 4.1, sont retenus dans cet article. On cherchera donc ici à minimiser :

- la somme des carrés des résidus,
- la somme des valeurs absolues des résidus,
- la médiane des carrés des résidus.

Dans ce qui suit, et en particulier dans le tableau 2, ces critères sont respectivement désignés par les symboles $\mathcal{O}_1$, $\mathcal{O}_2$, et $\mathcal{O}_3$.

Toute partition admissible c définit des moyennes ajustées $\underline{m}(c)$. Ces moyennes peuvent donc être comparées à celles obtenues dans les trois régressions sans contrainte effectuées au sens des trois critères d'optimisation $\mathcal{O}_j$ ($j \in \{1, 2, 3\}$). Par ailleurs, chaque partition admissible c peut être améliorée par application de l'algorithme de transfert piloté par les mêmes trois critères et donne naissance, à son tour, à trois partitions dérivées notées c', c'', et c'''. Il en résulte 12 comparaisons possibles, obtenues en croisant les 4 partitions admissibles c, c', c'', et c''' et les 3 régressions au sens des critères $\mathcal{O}_1$, $\mathcal{O}_2$, et $\mathcal{O}_3$. Chaque comparaison peut être quantifiée en calculant le rapport $\frac{u}{v}$ (exprimé en pourcentage) de la valeur u du critère d'optimisation atteinte sous contrainte à la valeur v obtenue sans contrainte. Ainsi, une valeur exacte de 100 % indique une situation dans laquelle la méthode de Mayer donne le même résultat qu'une régression ; une valeur de 150 %, indique une situation dans laquelle la méthode de Mayer dégrade la valeur du critère de régression à hauteur de 50 %.

Tous les calculs sont effectués à l'aide du logiciel statistique[14] R (cf. [R development Core Team, 2006] et [Venables et Ripley, 2002]). On compare l'ajustement constaté sur 12 partitions à celui obtenu par régression sans contrainte au sens des trois critères $\mathcal{O}_j$, $j \in \{1, 2, 3\}$, à minimiser. Les 12 partitions sont obtenues comme suit :

- Les quatre partitions initiales des données sont : celle, notée c_1, fournie par Mayer ; une partition, notée c_2, obtenue par application de la fonction *kmeans* du logiciel R en ne considérant que les variables explicatives (centrées et réduites) ; une partition, notée c_3, obtenue par application de la fonction *kmeans* du logiciel R en considérant l'ensemble des variables (réponse et explicatives, toutes centrées et réduites) ; enfin une partition, notée c_4, déduite d'un arbre de régression fourni par la fonction *rpart* de la bibliothèque *rpart*.

[14] Il s'agit de la version 2.3.1 (2006-06-01) d'un logiciel statistique gratuit et multi-plateformes. Il est téléchargeable à partir du site `http://cran.r-project.org/`. Le logiciel est complété par un certain nombre de bibliothèques (en anglais, *library*) de fonctions. Ces bibliothèques sont aussi téléchargeables et gratuites.

– Chaque partition initiale admissible c_i, $i \in \{1, \ldots, 4\}$, est optimisée par application de l'algorithme de transfert piloté successivement par l'un des trois critères d'optimisation retenus[15] ; les trois partitions admissibles construites à partir de la partition c_i sont notées c_i', c_i'', et c_i'''.

Les trois régressions sans contraintes qui correspondent à l'optimisation des trois critères cités sont estimées par les fonctions suivantes du logiciel R :

– Critère $\mathcal{O}_1$ (moindre somme des carrés des résidus) : fonction *glm* (cf. [McCullagh, Nelder, 1989]).
– Critère $\mathcal{O}_2$ (moindre somme des valeurs absolues des résidus) : fonction *qr* de la bibliothèque *quantreg* (cf. [Koenker, 2006]).
– Critère $\mathcal{O}_3$ (moindre médiane des carrés des résidus) : fonction *lqs* de la bibliothèque *MASS* [Venables et Ripley, 2002]).

Les résultats des 48 comparaisons effectuées sont rassemblés dans le Tableau 2. Ce dernier est composé de 4 blocs correspondant aux 4 initialisations effectuées. Comme, on pouvait s'y attendre, les indices d'optimalité relatifs à un même choix de critère d'optimalité (chiffres en gras dans le Tableau 2) montrent une bonne performance de l'algorithme de transfert : chiffres proches de 100 % pour les critères $\mathcal{O}_1$ et $\mathcal{O}_2$, de l'ordre de 110 à 140 % pour le critère $\mathcal{O}_3$. Toutefois, l'optimisation sous contrainte par la méthode des transferts est plus efficace pour les critères $\mathcal{O}_1$ ou $\mathcal{O}_2$ que $\mathcal{O}_3$. On notera, en passant, la robustesse vis-à-vis du choix du critère d'optimalité de la partition considérée par Mayer (chiffres en italiques dans le Tableau 2).

6. CONCLUSION

De nos jours, la régression linéaire sous contraintes fait partie de la boîte à outils de la statistique appliquée. Les contraintes portent le plus souvent sur les coefficients de la régression et se traduisent par des équations ou des inéquations linéaires et non-linéaires. La fonction objectif la plus classique est sans doute le critère des moindres carrés (cf. par exemple [Cazes, 1975]). Naturellement, d'autres fonctions objectifs que la somme des carrés des résidus de l'ajustement peuvent être retenues, les moindres valeurs absolues par exemple. Mais dans cet article, les contraintes sont de type très particulier : les solutions admissibles sont associées à des partitions.

La présentation combinée de l'heuristique de Mayer et de critères variés d'ajustement fournit ainsi des exemples inédits de régression sous contraintes. La formulation générale du problème recouvre des situations très variées : c'est d'abord Mayer (1750) qui recherchait une solution admissible sans avoir explicité de critère d'optimisation ; c'est ensuite Legendre (1805) qui explicitait un critère d'optimisation, les moindres carrés, mais n'imposait aucune contrainte ; mais il est des situations intermédiaires où l'algorithme des transferts de Régnier permet d'obtenir, pour un certain nombre de critères d'optimisation, une solution associée à une partition.

[15] Vu la petite taille de l'ensemble des données, on a effectué à chaque étape le transfert apportant la meilleure amélioration, tout en sachant que ce choix « local » ne conduit pas *in fine* à une solution nécessairement « meilleure » que celle que l'on aurait obtenue en choisissant systématiquement la première solution améliorante rencontrée.

TABLEAU 2. Exemple de Mayer. Comparaison des régressions de Mayer associées à différentes partitions et de régressions correspondant à trois critères classiques d'ajustement : moindres carrés des résidus $\mathcal{O}_1$, moindres valeurs absolues des résidus $\mathcal{O}_2$, moindre médiane des carrés des résidus $\mathcal{O}_3$. Les valeurs de l'indice d'optimalité sont arrondies à l'entier près. Cet indice, supérieur ou égal à 100 %, est d'autant plus grand que la contrainte de représentation par une partition est pénalisante.

Indice d'optimalité (en %)

	critère d'optimisation		
	$\mathcal{O}_1$	$\mathcal{O}_2$	$\mathcal{O}_3$
Partition de Mayer			
- c_1 : initiale	*101*	*103*	*167*
- c_1' : optimisée $\mathcal{O}_1$	**100**	105	194
- c_1'' : optimisée $\mathcal{O}_2$	113	**101**	163
- c_1''' : optimisée $\mathcal{O}_3$	115	103	**121**
Partition des K-moyennes sur variables explicatives seules			
- c_2 : initiale	100	106	202
- c_2' : optimisée $\mathcal{O}_1$	**100**	105	195
- c_2'' : optimisée $\mathcal{O}_2$	113	**101**	163
- c_2''' : optimisée $\mathcal{O}_3$	121	106	**110**
Partition des K-moyennes sur variables explicatives et réponse			
- c_3 : initiale	102	111	226
- c_3' : optimisée $\mathcal{O}_1$	**100**	105	185
- c_3'' : optimisée $\mathcal{O}_2$	110	**101**	150
- c_3''' : optimisée $\mathcal{O}_3$	124	109	**107**
Partition de l'arbre de régression			
- c_4 : initiale	120	129	281
- c_4' : optimisée $\mathcal{O}_1$	**100**	105	199
- c_4'' : optimisée $\mathcal{O}_2$	110	**101**	151
- c_4''' : optimisée $\mathcal{O}_3$	112	106	**141**

En étudiant plus avant le problème ainsi posé dans sa généralité, il est naturel d'explorer deux voies : celle de la relaxation des contraintes ou celle de l'amélioration de l'algorithme de résolution.

1. Relaxation des contraintes :
   - Le plus simple est de recourir à un recouvrement de l'ensemble des observations (partition à classes empiétantes) plutôt qu'à une partition. La matrice $\mathbf{C}$ est toujours à valeurs dans $\{0, 1\}$ mais certains totaux lignes sont strictement supérieurs à 1.
   - Plus élaboré consiste à s'affranchir de la contrainte de recouvrement tout en conservant la simplicité des calculs. La matrice $\mathbf{C}$ est alors à valeurs dans $\{-1, 0, 1\}$. Cette démarche a été considérée par Laplace[16] en 1788 (cf. [Hald,

[16]Pierre-Simon Laplace (1749-1827) est un mathématicien, astronome, physicien et philosophe français.

2007, p. 49 ; ou Farebrother, 1998, p. 29-31]). Ce dernier ne donne pas de méthode générale de construction de la matrice **C**.

Dans les deux cas, l'algorithme de transfert proposé dans cet article doit être adapté.

2. Efficacité de l'algorithme :
   L'analyse des données de l'exemple de Mayer montre que l'algorithme proposé dans cet article permet de construire des partitions admissibles qui sont assez performantes ; mais il s'agit là d'un petit ensemble de données ($n = 27$, p=3) ! Pour de très grands ensembles de données, il ne fait pas de doute que l'algorithme de transfert soit trop lent. Des algorithmes plus performants devraient alors être utilisés. Un arbitre suggérait l'emploi de métaheuristiques d'optimisation telle que la recherche tabou [Dreo, *et al.*, 2003].

Enfin, la démarche générale d'optimisation sous contrainte de partition pourrait être étendue à d'autres outils de la statistique, l'analyse en composantes principales (ACP) par exemple. Dans cette approche, un sous-espace principal d'ordre p serait alors approché par le sous-espace engendré par les centres de $p+1$ classes. L'interprétation de l'ACP pourrait être ainsi facilitée en rapprochant de façon assez naturelle les notions de dimension de représentation et de classificabilité.

Remerciements. Cet article est le développement d'un exposé présenté lors du « Colloque international de statistique appliquée pour le développement en Afrique, SADA'07 » (Cotonou, Bénin) et des « Rencontres de la Société française de statistique, SFC'07 » (Paris, France) tenus tous deux en 2007. L'auteur voudrait exprimer sa gratitude aux arbitres de ce journal pour leur lecture parfois stricte mais toujours juste d'une première version de cet article. La présente version leur doit beaucoup ; les erreurs ou incompréhensions qui demeurent sont de ma responsabilité.

BIBLIOGRAPHIE

BOCK H.-H., DIDAY E., *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*, Heidelberg, Springer, 2000, [2nd edition].

CAZES P., « Protection de la régression par utilisation de contraintes linéaires et non linéaires », *Revue de statistique appliquée*, tome 23(3), 1975, p. 37-57.

DRÉO J., PÉTROWSKI A., SIARRY P., TAILLARD É., *Métaheuristiques pour l'optimisation difficile*, Paris, Eyrolles, 2003.

FALGUEROLLES A. (de), PINCHON D., « Une commémoration du bicentenaire de la publication en 1805 (et 1806) de la méthode des moindres quarrés par Adrien Marie Legendre », *Journal de la Société française de statistique*, vol. 147(2), 2006, p. 81-105.

FAREBROTHER R. W., *Fitting linear relationships, a history of the calculus of observations (1750-1900)*, New York, Springer, 1998.

HALD A., *Sources and studies in the history of mathematics and physical sciences*, New York, Springer, 2007.

KOENKER R., *quantreg : Quantile Regression*, R package version 4.01, 2006, [`http://cran.r-project.org/`].

LEGENDRE A. M., *Nouvelle méthode pour la détermination des orbites des comètes*, Paris, Courcier, 1805.

MAYER J. T., "Abhandlung über die Umwälzung des Monds um seine Axen und die scheinbare Bewegung der Mondsflecten", *Kosmographische Nachrichten und Sammlungen auf das Jahr 1748*, 1, 1750, p. 52-183.

MCCULLAGH P., NELDER J. A., *Generalized linear models*, London, Chapman and Hall, 1989, [2nd edition].

"R Development Core Team", *R : A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2006, [`http://cran.r-project.org/`].

RÉGNIER S., « Sur quelques aspects mathématiques de problèmes de classification automatique », *I.C.C. Bulletin* 4, 1965, p. 175-191, et *Mathématiques et Sciences humaines* 82, 1983, p. 13-29.

STIGLER S. M., *The history of statistics : the measurement of uncertainty before 1900*, Cambridge (Massachusetts), The Belknap Press of Harvard University, 1986.

STIGLER S. M., *Statistics on the table, the history of statistical concepts and methods*, Cambridge (Massachusetts), Harvard University Press, 1999.

TASSI PH., *Méthodes statistiques*, Paris, Economica, 1989, [2^{e} édition].

TIBSHIRANI R., "Regression shrinkage and selection via the lasso", *J. Royal Statist. Soc. (Series B)* 58, 1996, p. 267-288.

VENABLES W. N., RIPLEY B. D., *Modern applied statistics with S*, New York, Springer, 2002, [4th edition].