LES CAHIERS DE L'ANALYSE DES DONNÉES

J.-P. BENZÉCRI
F. BENZÉCRI

Codage linéaire par morceaux et équation personnelle

Les cahiers de l'analyse des données, tome 14, n° 3 (1989), p. 331-336

<http://www.numdam.org/item?id=CAD_1989__14_3_331_0>

Les Cahiers de l'Analyse des Données
Vol XIV - 1989 - n°3 - pp. 331-336.

CODAGE LINÉAIRE PAR MORCEAUX ET ÉQUATION PERSONNELLE

[ÉQ. PERS.]

J.-P. & F. BENZÉCRI

1 Analyse de tableaux de notes et équation personnelle

Il est fréquent que se rencontre dans un questionnaire un bloc de questions de même format auxquelles le sujet doit répondre en notant sur un intervalle donné (par exemple de 0 à 10; en valeurs entières ou par degrés continus) son degré d'approbation pour une suite de phrases, son estime pour plusieurs hommes politiques, la qualité d'un ensemble de plats, etc...

Il est bien connu que les sujets, en présence d'une telle échelle présentant plusieurs degrés, ont un comportement qui n'est que partiellement déterminé par leur opinion sur le thème de la question. Indépendamment de ce thème, certains tendent à donner des réponses extrêmes, soit favorables, soit défavorables; d'autres se refusent à quitter le centre de l'échelle.

Nous savons bien que plus les sujets interrogés ont de liberté dans leur réponse, plus celle-ci est difficile à interpréter. D'un autre côté, un format trop rigide (tel que 'pour' ou 'contre', deux modalités) met le sujet si mal à l'aise qu'on ne peut se fier à ce qu'il a dit. Le statisticien doit s'accommoder des données recueillies, tout en s'appliquant à en aviver, en quelque sorte, la couleur, par le codage.

Le codage de chaque note j suivant trois modalités {j+, j=, j-}, avec une formule, dite *équation personnelle*, propre à chaque sujet, a déjà servi dans plusieurs études, (cf., notamment [ERGO. RÉGLAGES], [POLIT. GREC], in *CAD*, Vol XIII, n°2, 1988; et [NOTES MOTS], in *CAD*, Vol XIV, n°1, 1989).

En bref, on recadre entre -1 et +1 l'ensemble des notes attribuées par un même sujet *i*, en en calculant la moyenne, le Max et le min. Les notes sont d'abord centrées en en retranchant la moyenne moy. Puis toute note >0 est divisée par (Max - moy); toute note <0 est divisée par (moy - min); ainsi les notes données par le sujet *i* varient de -1 à +1. Soit maintenant k(i,j) une note recadrée; on la code sur 3 modalités par la formule:

Les cahiers de l'analyse des données – 0339-3097/89/03 331 6/$ 2.60/

si (k(i,j)≤0) alors début k(i,j+):=0; k(i,j=):=1+k(i,j); k(i,j-):=-k(i,j) fin
sinon début k(i,j+):=k(i,j); k(i,j=):=1-k(i,j); k(i,j-):=0 fin.

On reconnaît sans peine, dans ce codage, un principe barycentrique (cf. [CODAGE LIN.], in *CAD*, Vol XIV, n°2, 1989).

Si, maintenant, on admet que les notes ainsi recadrées ont une signification univoque, (l'effet de l'équation personnelle ayant été bien compris), on peut reprendre sur les colonnes ce qui a été fait sur les lignes, soit:

1° Calculer la moyenne myj, le maximum Mxj et le minimum mnj des notes, (notes recadrées entre -1 et +1), figurant dans la colonne j;

2° Recadrer la colonne j suivant la formule:

si (k(i, j) > myj) alors k(i, j) := (k(i, j) - myj) / (Mxj - myj)
sinon k(i, j) := (k(i, j) - myj) / (myj - mnj) ;

3° Comme précédemment, coder la note recadrée suivant 3 modalités qu'on pourra noter j<, j≈ et j> afin de les distinguer de celles précédemment introduites.

L'intérêt de ce double recadrage (déjà évoqué dans [NOTES MOTS], *in fine*) est de produire des notes j> et j< équilibrées, qu'il s'agisse d'une entité *j* (phrase ou personne) généralement bien notée ou mal notée; (alors que la modalité j- est d'autant plus lourde que *j* est plus mal notée par la majorité des sujets). En effet, ici, le point moyen utilisé pour le deuxième recadrage, est adapté chaque fois à *j*. Il s'agit, ici encore, d'un codage barycentrique.

On peut effectuer les deux codages par l'option 'Q' (barycentrique) du programme 'zrang' (cf. [CODAGE LIN.]) à condition d'avoir préalablement effectué, par un programme approprié, le cadrage des notes de chaque ligne *i* entre -1 et +1 suivant l'équation personnelle. Pour le recadrage simple, on prend pour toute variable *j* les 3 valeurs pivot {-1, 0, +1}; tandis que pour le double recadrage les valeurs pivot sont {mnj, myj, Mxj}, calculées sur chaque colonne *j* comme on l'a expliqué ci-dessus.

Il faut cependant prendre garde que les questionnaires sont rarement remplis sans qu'y manquent certaines réponses, que l'on convient de remplacer par un nombre tel que 0 ou 9; ou encore, en format réel, e+25. Le calcul de l'équation personnelle ne peut guère être fait que sur un sous-tableau ne comportant aucune donnée manquante; tableau construit, de préférence, en éliminant celles des lignes (individus) et des colonnes (questions) qui comportent le plus de lacunes.

L'objet du présent article est de montrer, sur un exemple précis, par quelle suite d'opérations on peut, à l'aide du logiciel Mac SAIF, tel qu'il a été présenté dans *CAD* Vol XIV n°1, et de deux nouveaux programmes 'manq' et 'pers',

extraire d'un questionnaire donné un tableau de notes sans lacune et analyser celui-ci après simple ou double recadrage suivant l'équation personnelle.

Nous distinguerons trois étapes successives: le choix du tableau de notes sans lacune (§2); le codage suivant l'équation personnelle (§3); l'adjonction de modalités supplémentaires provenant de questions du même questionnaire mais auxquelles les réponses ne sont pas sous forme de notes (§4).

2 Du questionnaire au tableau de notes sans lacune

L'exemple choisi est celui d'un questionnaire proposé à 258 étudiants de l'Université Concordia de Montréal et relatif au mémoire de recherche de la maîtrise en administration des affaires (cf.[QUEST. MÉM. RECH.], in *CAD*, Vol XIV, n°3, 1989). Le questionnaire comprend 46 questions, dont 34 (q. 2 à 35) proposent une phrase à laquelle on doit répondre sur une échelle d'approbation à 5 degrés allant de 1 (désapprobation totale) à 5 (approbation totale).

Les données constituent un tableau 'mba' (plus précisément 'D:mba', si 'D' est le disque, ou dossier, utilisé), de format texte, avec, conformément aux spécifications de Mac SAIF, une ligne de titre suivie du nombre de colonnes (46) et des identificateurs de celles-ci (qui sont simplement formés de la lettre *V* suivie d'un numéro de 1 à 46), puis des lignes afférentes à chacun des individus *i*:; avec un sigle (qui est un simple numéro) et 46 nombres entiers qui désignent les modalités de réponses choisies par *i*. Pour les questions 2 à 35 (phrases) il y a 6 modalités de réponse possibles: d'une part les nombres de 1 à 5 (niveaux d'approbation) et d'autre part le *zéro* (omission, non concerné).

Le tableau 'mba' est d'abord copié par 'zrang' (option 'C') en un tableau de format 'ww' (réels): 'mbaww'. Puis, de ce tableau, on extrait par 'soustab' le sous-tableau formé des colonnes 2 à 35 contenant les modalités de réponse aux phrases: on a ainsi un tableau 'D:mba1ww', (258 × 34).

De 'D:mba1ww' on désire extraire un tableau sans lacune; c'est-à-dire, avec les notations adoptées, sans chiffre 0. À cette fin, on ouvre le programme 'manq'. S'affiche la phrase:

```
ce programme recense les données manquantes
```

suivie de la question:

```
le fichier ds données est
```

à quoi l'on répond: 'D:mba1' (sans le suffixe'ww); et le tableau s'affiche au fur et à mesure de son entrée en mémoire centrale. Puis vient la question:

```
la valeur attribuee aux donnees manquantes est
```

à laquelle on répond: 0 (zéro).

Le programme crée alors un listage 'D:mba1mqx', tout en affichant quelques indications sur les manquants sous la forme (e.g.):

```
i = 17 manque(nt) 23
```

pour indiquer qu'il y a 23 lacunes à la ligne 17; et de même, ensuite, pour les colonnes.

Le listage 'D:mba1mqx', se compose de deux parties de même format, la première afférente aux lignes , la deuxième aux colonnes. La première partie donne d'abord, sous le titre "dénombrement ligne par ligne", une suite d'alinéas dont chacun concerne une ligne comportant effectivement des manques; avec le numéro de la ligne ('ligne xx') suivi des numéros des colonnes où sont les lacunes. Vient ensuite une "liste des numéros des lignes sans donnée manquante". La deuxième partie, placée sous le titre "dénombrement colonne par colonne", a même structure que la première.

En consultant 'D:mbamqx', on constate que 70 lignes et 3 colonnes seulement sont sans lacune. Les utilisateurs ont choisi d'éliminer d'abord les colonnes comptant le plus de lacunes; plus précisément les 9 colonnes ayant plus de 15 lacunes. On constitue donc une liste des numéros des 25 colonnes à garder dans un tableau 'D:mba11ww' qui sera créé par le programme 'soustab': c'est le fichier texte 'D:mba11ensj' ci-dessous (cf. [NOTE CRé. TAB.], §5)

```
COLONNES connservees pour 1 analyse
          3  4  5            9
10 11 12    14 15    17 18 19
20    22 23 24 25 26 27    29
30 31 32 33 34
```

On ouvre alors 'soustab' en demandant d'extraire de 'D:mba1' un sous-tableau dont le nom aura le suffixe '1'; les 25 colonnes à garder étant spécifiées par le fichier 'ensj'; et les lignes demandées en dialogue, comme formant un seul bloc [1..258] (i.e., toutes les lignes).

Le tableau 'D:mba11ww', ainsi créé, est lui-même soumis à 'manq'. On constate qu'il y a 183 lignes sans lacunes. On prend la liste de celles-ci, opportunément donnée sur le listage 'D:mba11mqx', pour constituer un fichier texte 'D:mba111ensi'. On ouvre une nouvelle fois 'soustab', en demandant d'extraire de 'D:mba11' un sous-tableau dont le nom aura le suffixe '1'; les colonnes étant toutes à garder (bloc [1..25]) et les lignes à garder étant spécifiées par le fichier 'ensi'.

Pour plus de sûreté, on peut vérifier que le tableau 'D:mba111ww', (183 × 25), ainsi créé ne contient pas de lacune en le soumettant à 'manq'. En effet, dans 'D:mba111mqx', il n'y a, après le titre "dénombrement ligne par ligne", aucun alinéa recensant des lacunes, mais seulement la "liste des numéros des lignes sans données manquantes", qui comprend tous les nombres de 1 à 183; et semblablement pour les colonnes.

3 Recadrage des notes entre -1 et +1 suivant l'équation personnelle

Le tableau 'D:mba111ww' ne contient que des notes de 1 à 5, effectivement attribuées par les 183 sujets aux 25 phrases. Afin de recadrer ces notes, on ouvre le programme 'pers'. S'affiche le titre:

```
programme pour l'équation personnelle
```

et, après la demande du nom de base du tableau à traiter (à quoi l'utilisateur répond 'D:mba111', sans le suffixe 'ww'), le tableau entre en mémoire centrale et 'per' crée trois fichiers:

'D:mba111§ww', qui est le tableau 183 × 25 des notes recadrées entre -1 et 1, ligne par ligne, suivant l'équation personnelle propre à chaque sujet;

'D:mba111§Dcodx', qui est un fichier texte dont le format est expliqué dans [NOT. CRÉ. TAB.], §2.2.3: plus précisément, il s'agit d'un tableau définissant pour chacune des variables, dont les sigles sont présentement de la forme 'Vx', des modalités et pivots qui sont toujours de la même forme, soit:

```
Vx a 3 modalites dont les sigles et pivots sont
Vx-   Vx=   Vx+
-1.000        0.000       1.000
```

Il faut seulement noter que si le sigle d'une variable comporte 4 caractères, et non 2 ou 3 comme ici, le dernier est supprimé avant d'ajouter '-', '=' ou '+' pour créer les sigles des modalités.

'D:mba111$Dcodx', enfin, est aussi un fichier 'Dcodx'; mais il est destiné à effectuer un double recadrage; les valeurs pivot ne sont donc pas les mêmes pour toutes les variables; elles sont calculées respectivement comme minimum, moyenne et Maximum (cf. *supra*, mnj, myj, Mxj) des colonnes correspondantes. De plus, afin d'éviter toute confusion, les sigles sont différents de ceux de '§Dcodx': '<' et '>' remplacent '-' et '+'.

Le programme 'pers' affiche à l'écran, au cours du calcul, les valeurs pivot des fichiers qu'il crée.

Pour obtenir le tableau 183 × 75, avec 3 modalités {-, =, +} par note recadrée une fois suivant l'équation personnelle, il suffit de soumettre à 'zrang' le tableau 'D:mba111§ww', en choisissant {'D', 'P', 'Q'}; i.e. Découper suivant des bornes Préétablies pour créer un tableau barycentriQue (présentement, 3 colonnes par variables de base). On obtient ainsi 'D:mba111§Qww', (183 × 75), qui peut être soumis à l'analyse factorielle, puis à la classification (programmes 'qori' et 'CAH2').

Pour obtenir le tableau 183 × 75, avec 3 modalités {<, =, >}, (double recadrage), on crée une copie de 'D:mba111$ww' de 'D:mba111§ww' (ou, si

l'on préfère, on renomme le tableau 'D:mba111§ww'); et on soumet, de même, 'D:mba111$ww' à 'zrang' pour créer le tableau 'D:mba111$Qww', (183 × 75); lequel peut également être soumis à 'qori' puis à 'CAH2'.

4 Adjonction de modalités supplémentaires provenant d'autres questions

Parmi les questions qui ne sont pas en forme de phrases proposées à l'approbation des sujets, 6 ont été retenues par les auteurs comme susceptibles d'être adjointes en modalités supplémentaires. On décrira la suite des opérations effectuées à cet effet.

Du tableau 'D:mbaww', on extrait d'abord un tableau nommé 'D:mba2ww', ayant pour lignes celles-là mêmes retenues pour le tableau 'D:mba111ww'; et pour colonnes, celles auxquelles on s'intéresse (et dont les numéros initiaux se trouvent être {1, 36, 37, 44, 45, 46}. On utilise le programme 'soustab', en spécifiant les lignes à garder par une copie de 'D:mba111ensi', renommée 'D:mba2ensi'; tandis que pour spécifier les colonnes, il est aussi simple de procéder par dialogue que de créer un fichier 'ensj'.

Le tableau 'D:mba2ww' est codé sous forme binaire, (0,1), en un tableau 'D:mba2Bbzz', comme expliqué dans [NOT. CRÉ. TAB.], §2.2. On utilise 'zrang': on peut soit procéder exclusivement par dialogue, soit observer d'abord les données et tenir compte du codage du questionnaire pour créer un fichier 'D:mba2Dcodx' sur un éditeur de texte. Sans sortir de 'zrang', on transpose 'D:mba2Bbzz' en un tableau 'D:mba2BbTww'.

Enfin, le tableau 'D:mba2BbTww', renommé 'D:mba111§Q2bww', peut être soumis à 'qorelsup', comme un tableau externe à adjoindre à 'D:mba111Q', en donnant pour sigle '2b'; ce qui est un nom convenable pour un tableau de colonnes supplémentaires (présenté transposé, comme on l'explique au §4.2 de [NOT. CRÉ. TAB.]).

On obtient un listage de facteurs, 'D:mba111§Q2bcorsutx'; et un fichier de coordonnées sur les axes factoriels 'D:mba111§Q2jbFacww'; ce fichier permet de placer les modalités supplémentaires sur les graphiques créés par 'planF' ou 'planX' (cf. [NOT. CORR. CAH.], §§2.2 & 2.3) en donnant '2jb' pour sigle de l'ensemble.

Il va sans dire qu'on procède de même pour adjoindre les mêmes éléments supplémentaires à l'analyse des notes deux fois recadrées; la seule différence étant que le tableau 'D:mba2BbTww' doit être renommé 'D:mba111$Q2bww', (avec '$' au lieu de '§').