LES CAHIERS DE L'ANALYSE DES DONNÉES

J. TCHOUANKAM

Un modèle d'analyse des réponses de sujets, de niveau différent, à des questions à double issue, de difficulté variable

Les cahiers de l'analyse des données, tome 17, n° 4 (1992), p. 395-402

<http://www.numdam.org/item?id=CAD_1992__17_4_395_0>

© Les cahiers de l'analyse des données, Dunod, 1992, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

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

Les Cahiers de l'Analyse des Données
Vol XVII - 1992 - n°4 - pp. 395-402.

UN MODÈLE D'ANALYSE DES RÉPONSES DE SUJETS, DE NIVEAU DIFFÉRENT, À DES QUESTIONS À DOUBLE ISSUE, DE DIFFICULTÉ VARIABLE

[DOUBL. DIFF. VAR.]

*J. TCHOUANKAM**

1 Origine du problème

On suppose qu'un questionnaire se compose exclusivement de questions fermées à deux issues, dont l'une est, en fait, la réponse vraie; et l'autre, une réponse fausse. Un sujet fournit la réponse vraie si ses connaissances le lui permettent; sinon, il répond, au hasard, équiprobablement, l'une ou l'autre réponse. En général, les questions porteront sur des domaines différents; un sujet sera à l'aise dans un domaine; un autre sujet dans un autre domaine.

Mais nous faisons, ici, l'hypothèse simplificatrice que sujets et questions sont caractérisés par une variable unique qu'on peut appeler le niveau, x(i), du sujet *i*; ou la difficulté, x(q), de la question *q*; le sujet *i* connaît la réponse exacte, qu'on peut convenir de noter q+, si son niveau, x(i), est supérieur ou égal à la difficulté x(q) de la question *q*; si, au contraire, $x(i) < x(q)$, il répond, avec une égale probabilité, l'une ou l'autre des deux réponses {q+, q-}.

Si I est l'ensemble des sujets et Q, l'ensemble des questions, il est d'usage de disposer les réponses suivant un tableau en (0, 1), $I \times J$, où $J = \{Q+ \cup Q-\}$. Notre propos est de considérer l'analyse du tableau $I \times J$, en nous bornant au cas limite où le nombre des sujets est assez élevé pour que joue la loi des grands nombres.

Partons du cas le plus simple où Q comprend *n* questions, toutes de la même difficulté, x+. Sous cette hypothèse, les sujets se répartissent en deux groupes: ceux dont le niveau est $\geq x+$, et qui fournissent à toutes les questions la réponse exacte, q+; et ceux de niveau $<x+$, qui répondent toujours au hasard. Si le nombre des sujets est élevé, de par la loi des grands nombres, il y a un même nombre de sujets de niveau $<x+$, pour fournir chacun des 2^n patrons de

(*) Étudiante en doctorat à l'Université Pierre et Marie Curie.

Les cahiers de l'analyse des données - 0339-3097/92/04 395 08/$ 2.80/© Gauthier-Villars

réponse possibles. Voici, par exemple ce qu'est le tableau des réponses dans le cas où il n'y a que deux questions, {q1, q2}, de niveau x+, et où il y a le même nombre de sujets de niveau ≥x+ et de niveau <x+:

cas modèle à deux questions				
4	q1+	q1-	q2+	q2-
i>	4	0	4	0
ia<	1	0	1	0
ib<	1	0	0	1
ic<	0	1	1	0
id<	0	1	0	1

cas modèle à deux questions				
4	q1+	q1-	q2+	q2-
i><	5	0	5	0
ib<	1	0	0	1
ic<	0	1	1	0
id<	0	1	0	1

à gauche, on suppose qu'il y a 4 sujets de niveau ≥x+, dont les réponses sont cumulées dans la ligne i>; et 4 sujets {ia<, ib<, ic<, id<}, de niveau <x+, fournissant chacun l'un des 4 patrons de réponse possibles; à droite, l'on a cumulé, en i><, les lignes {i>, ia<}, de même profil. Il est clair qu'avec un grand nombre de sujets, également répartis entre niveau ≥x+ et niveau <x+, on aura, aux fluctuations près, un tableau proportionnel au tableau de droite, pourvu que l'on cumule les lignes identiques.

Avec *n* questions de même niveau x+, et une fraction *m* des sujets ayant un niveau ≥x+, l'on aura, de même, un tableau type, à 2n colonnes et 2^n lignes; avec le poids ((1-m) + m.2^n) pour la première ligne, i><, correspondant au patron où toutes les réponses sont exactes; et le poids (1-m) pour chacune des (2^n-1) autres lignes, correspondant à un patron qui comporte au moins une erreur.

2 Description d'un modèle général

On peut maintenant décrire le modèle le plus général compatible avec les hypothèses que nous avons posées.

Les questions se répartissent en *p* niveaux distincts de difficulté:

$$x_1 < x_2 < x_3 < \ldots < x_h < \ldots < x_{p-1} < x_p \ ;$$

il y a, au niveau x_h, un ensemble Q_h de n_h questions; l'ensemble Q de toutes les questions a pour cardinal $n = \sum\{ n_h \mid h = 1, \ldots, p\}$.

Quant aux sujets, il suffit de donner les (p+1) probabilités, de total 1, $m_0, \ldots, m_h, \ldots, m_p$, qu'ils se trouvent dans chacun des (p+1) intervalles successifs $X_0, \ldots, X_h, \ldots, X_p$:

$X_0 =]-\infty, x_1[$; $X_1 = [x_1, x_2[$; ... ; $X_h = [x_h, x_{h+1}[$; ... ; $X_p = [x_p, \infty[$;

Ainsi, l'ensemble I des 2^n patrons de réponse est réparti en (p+1) sous-ensembles emboîtés $I_p, I_{p-1}, \ldots, I_h, \ldots, I_1, I_0$, caractérisés comme suit:

I_p comprend l'unique patron dont toutes les réponses sont exactes;

$I_{p\,1}$ comprend les patrons donnant des réponses exactes de Q_1 à $Q_{p\,1}$ mais comportant au moins une erreur dans le bloc Q_p ;

I_h comprend les patrons donnant des réponses exactes de Q_1 à Q_h mais comportant au moins une erreur dans Q_{h+1} ;

I_0 comprend les patrons qui ont au moins une erreur dans Q_1 .

Dans le calcul des poids des patrons, il est commode de normaliser ceux-ci en sorte qu'ils aient le total 2^n ; il vient alors:

$$\forall\ i \in I_0 : \text{poids} = m_0 ;$$

$$\forall\ i \in I_1 : \text{poids} = m_0 + m_1 \,.\, 2^{n1} ;$$

$$\forall\ i \in I_2 : \text{poids} = m_0 + m_1 \,.\, 2^{n1} + m_2 \,.\, 2^{n1+n2} ;$$

$$\forall\ i \in I_h : \text{poids} = m_0 + m_1 \,.\, 2^{n1} + \ldots + m_h \,.\, 2^{n1+n2+\ldots+nh} ;$$

$$\forall\ i \in I_p : \text{poids} = m_0 + m_1 \,.\, 2^{n1} + \ldots + m_p \,.\, 2^{n1+n2+\ldots+np} ;$$

pour vérifier que le total des poids est bien 2^n, il suffit de remarquer que chacun des 2^n patrons reçoit la masse m_0 ; que les $(2^n / 2^{n1})$ patrons sans erreur dans Q_1 mais quelconques ailleurs, reçoivent en outre la masse $m_1.2^{n1}$; que les $(2^n / 2^{n1+n2})$ sans erreur dans Q_1 et Q_2 mais quelconque ailleurs, reçoivent encore la masse $m_2.2^{n1+n2}$; etc... ; d'où, finalement, puisque $1 = m_0 + m_1 + \ldots$, la masse totale 2^n.

Finalement, il apparaît que le tableau I × J, afférent au modèle général, dépend des paramètres suivants:

un entier *p* strictement positif, le nombre des niveaux de difficulté distingués pour les questions: dans le cas, très simple, considéré au §1, ce nombre est 1;

une suite d'entiers, strictement positifs, $\{n_1, \ldots, n_h, \ldots, n_p\}$, donnant le nombre des questions à chaque niveau;

une suite de réels, strictement positifs, $\{m_0, \ldots, m_h, \ldots, m_p\}$, de somme 1, donnant la répartition des niveaux de compétence des sujets relativement aux niveaux de difficulté des questions.

Il est utile pour la suite de préciser également la forme du tableau de BURT, $J \times J$, afférent au modèle général. Il suffit de calculer le bloc carré (2×2), $\{q+, q-\}\times\{q'+,q'-\}$, croisant les modalités de deux questions $\{q, q'\}$. Ici, on normalisera en attribuant la valeur 1 au total des masses des sujets (ou des patrons). Nous considérerons successivement les 3 cas possibles: q=q', q et q' sont distinctes mais de même niveau, q et q' ont des niveaux distincts.

Pour une seule question q, on a:

$$k(q+, q+) = \mu \quad ; \quad k(q+, q-) = k(q-, q+) = 0 \quad ; \quad k(q-, q-) = 1-\mu \quad ;$$

où μ désigne la probabilité qu'un sujet réponde correctement à la question q: il est aisé de calculer $(1-\mu)$, probabilité d'erreur, qui n'est autre que la moitié du poids des sujets dont le niveau est strictement inférieur à la difficulté de la question. Pour $q \in Q_h$ on a donc:

$$1 - \mu_h = (1/2) \cdot (m_0 + \ldots + m_{h-1}) \quad ;$$

$$\mu_h = 1 - ((1/2) \cdot (m_0 + \ldots + m_{h-1})) \quad .$$

Pour deux questions, $\{q, q'\}$, de même niveau, x_h, on a de même:

$$k(q+, q'+) = 1 - ((3/4) \cdot (m_0 + \ldots + m_{h-1})) = (1/2) \cdot (3.\mu_h - 1) \quad ;$$

$$k(q+, q'-) = k(q-, q'+) = k(q-, q'-) = (1/4) \cdot (m_0 + \ldots + m_{h-1}) = (1/2) \cdot (1 - \mu_h) \; ;$$

en effet, en bref, les réponses inexactes, à l'une ou l'autre des questions, ne peuvent provenir que de sujets dont le niveau est insuffisant; et ceux-ci se répartissent également entre les 4 combinaisons de réponses possibles.

Relativement à deux questions, $\{q, q'\}$, de niveaux différents, $x_h<x_{h'}$, les sujets doivent, selon leur niveau, niv, être répartis en trois groupes, dont nous précisons les poids (dont la somme est 1):

$$\text{niv} < x_h \quad : \quad \mu_a = (m_0 + \ldots + m_{h\,1}) \; ; \quad (a) \; ;$$

$$x_h \le \text{niv} < x_{h'} \quad : \quad \mu_b = (m_h + \ldots + m_{h'-1}) \; ; \quad (b) \; ;$$

$$x_{h'} \le \text{niv} \quad : \quad \mu_c = (m_{h'} + \ldots + m_p) \; ; \quad (c) \; ;$$

les sujets du niveau (c) ne peuvent répondre que $\{q+, q'+\}$; ceux du niveau (b) se partagent également entre $\{q+, q'+\}$ et $\{q+, q'-\}$; ceux du niveau (a), totalement incompétents se partagent en quarts; d'où:

$$k(q+, q'+) = (1/4)\,\mu_a + (1/2)\,\mu_b + \mu_c \quad ; \quad k(q-, q'-) = (1/4)\,\mu_a \; ;$$

$$k(q+, q'-) = k(q+, q'-) = (1/4)\,\mu_a + (1/2)\,\mu_b \; ;$$

ainsi est complètement décrit le tableau de BURT.

3 Propriétés des facteurs issus du modèle général

Le tableau I × J, rentre dans le cadre de modèles classiques; ce qui permet d'en réduire l'analyse à la diagonalisation d'une matrice carrée symétrique de rang p; l'analyse s'achevant ensuite par des calculs simples ne comportant pas de processus itératif.

D'abord, le tableau admet un groupe de symétrie, qui n'est autre que le groupe **G**, de cardinal $\prod\{n_h! \mid h=1, \ldots, p\}$, des permutations g de Q laissant globalement invariant chacun des sous-ensembles Q_h. En effet, en bref, une telle permutation induit, d'une part, une permutation de J; et, d'autre part, une permutation de I qui échange entre eux des patrons ayant le même poids; donc conserve le tableau I × J.

L'analyse d'un tableau invariant par permutation simultanée de ses lignes et de ses colonnes est exposé dans ENS2, *Abrégé Théorique, Études de cas modèle*,IV n°0, §1.2. Notons de la même lettre g les permutations simultanées des deux ensembles I et J: il est clair que, si (F, G) est un couple de facteurs associés, ($F \circ g$, $G \circ g$) est également un couple de facteurs associés, relatif à la même valeur propre.

De plus, le tableau I × J offre un exemple très simple de tableau de notes dédoublé: en effet, pour toute question q et tout sujet i, la somme des deux notes k(i, q+) et k(i, q-) vaut 1. Dans ce cas, F. NAKHLÉ, a montré (cf. *CAD*, Vol I, n°3, pp. 243 sqq, 1976; ou ENS2, V n°1) que les facteurs sur J sont déterminés par diagonalisation d'une matrice symétrique S, Q × Q, i.e., de rang n, selon les notations du §2.

Pour la matrice S, la propriété de symétrie est qu'elle est invariante par permutation simultanée de l'ensemble Q des lignes et de l'ensemble Q des colonnes par une même permutation g de **G**; donc, si f désigne un vecteur propre de S, considéré comme fonction sur Q, $f \circ g$ est aussi un vecteur propre, relatif à la même valeur propre; et, plus généralement, si k(g) est une fonction quelconque, à valeur réelle, sur le groupe **G** la combinaison linéaire:

$$\Sigma\{k(g)\, f \circ g \mid g \in \mathbf{G}\},$$

est un vecteur propre, f_k, relatif à la même valeur propre que f.

Pour diagonaliser S, nous suivrons deux voies différentes; d'une part, considérer l'action du groupe **G** et de ses sous-groupes; d'autre part, préciser la structure des blocs de S. La deuxième voie est plus simple, mais la première passe par des constructions géométriques plus générales. Ayant ainsi découvert la structure de l'ensemble des facteurs, il nous apparaîtra plus simple d'effectuer des calculs directement sur le tableau de BURT lui-même; et nous donnerons un aperçu de ces calculs.

Notons E l'espace des fonctions sur Q muni de la métrique euclidienne usuelle, somme des carrés des composantes; E_0, le sous-espace de E, ensemble des fonctions constantes sur chacun des Q_h ; E_0 est de dimension p. Notons E_h, pour h = 1, ..., p, le sous-espace de E, ensemble des fonctions de moyenne nulle sur Q_h, et nulles sur chacun des autres $Q_{h'}$; E_h est de dimension (n_h-1). Il est clair que E est somme directe orthogonale des E_h:

$$E = \oplus \{E_h \mid h = 0, 1, 2, \ldots, p\} \ ;$$

de plus, chacun des E_h est sous-espace invariant par S.

Afin de démontrer cette propriété d'invariance, on caractérisera les sous-espaces E_h en terme de permutations et d'orthogonalité. Ici, il est commode de poser quelques notations: $Q_h!$, E(h), N(h).

$Q_h!$ est le groupe des $n_h!$ permutations de Q qui, d'une part, laissent invariant Q_h dans son ensemble, et, d'autre part, laissent invariant chacun des $(n - n_h)$ éléments de $Q - Q_h$.

E(h), (h≠0), désigne l'ensemble des fonctions de E constantes sur Q_h et quelconques ailleurs. Le sous-espace E(h) est caractérisé par l'équation:

$$\forall f \in E(h) : f = (1/(n_h!)) \, . \, \Sigma\{f \circ g \mid g \in Q_h! \} \ ;$$

cette caractérisation commute avec l'action de S sur f, parce que l'action de S commute avec la composition avec toute permutation du groupe **G**; donc l'image de E(h) par S est incluse dans E(h). Ceci posé, il est clair que E_0 est envoyé dans lui-même par S, comme intersection des p sous-espaces E(h) qui jouissent chacun de cette propriété.

N(h), (h≠0), est le sous-espace supplémentaire orthogonal de E(h) dans E: N(h) n'est autre que l'ensemble des fonctions de moyenne nulle sur Q_h , et nulles sur le complémentaire $Q - Q_h$: donc $N(h) = E_h$; N(h) est envoyé dans lui-même par l'application symétrique S qui conserve E(h) ; en effet, soit $f \in$ N(h) et $f' \in$ E(h): de par la symétrie de S on a l'égalité de produits scalaires:

$$\langle Sf, f' \rangle = \langle f, Sf' \rangle = 0 \ ;$$

où la valeur nulle résulte de ce que (cf. *supra*) $Sf' \in$ E(h); donc Sf, orthogonal à tout vecteur de E(h), est dans N(h).

Pour diagonaliser S, il suffit de considérer sa restriction à chacun des sous–espaces E_h ; pour E_0, on doit diagonaliser la matrice (p × p) obtenue à partir de S, en cumulant chacun des blocs $(Q_h \times Q_{h'})$ et divisant le résultat du cumul par $\sqrt{(n_h.n_{h'})}$; pour les autres E_h , (h≠0), le problème est encore plus simple, car il apparaît que, sur un tel sous–espace, S agit comme une homothétie; i.e., tout vecteur est vecteur propre, relatif à la même valeur propre.

En effet, soit $f \in E_h$, vecteur propre de S relatif à une valeur propre λ, ($f \neq 0$), k(g) une fonction quelconque sur $Q_h!$; on a dit que la combinaison linéaire:

$$f_k = \Sigma\{k(g)\, f \circ g \mid g \in Q_h!\ \} \in E_h \quad ,$$

est vecteur propre de S, pour la même valeur propre λ ; or on va montrer qu'avec une fonction k(g) convenable, on peut obtenir pour f_k tout vecteur de E_h , i.e. toute fonction de moyenne nulle sur Q_h.

Il est clair qu'il suffit de montrer qu'on peut obtenir un vecteur f_k non nul seulement en deux points, {q, q'} de Q_h: car alors, par action du groupe $Q_h!$, on obtient toute fonction sur Q_h nulle partout sauf en deux points déterminés où elle vaut {+1, -1}; donc, par combinaison linéaire, toute fonction de E_h. La démonstration repose sur le fait que si f est une fonction de moyenne nulle sur Q_h, prenant des valeurs non nulles en r points de Q_h, avec r>2, on peut trouver un nombre x>0 et une permutation g, échangeant seulement deux éléments {q, q'} de Q_h , de telle sorte que (f+ x.$f \circ g$) soit $\neq 0$ et prenne des valeurs non nulles en au plus (r-1) points de Q_h. Pour cela, il suffit de prendre pour {q, q'} deux points où f a des valeurs $\neq 0$ et de signes opposés; et de poser x = $|f(q)/f(q')|$; ainsi, (f+ x.$f \circ g$) s'annule en q; et il est, en dehors de {q, q'}, le multiple de f par (1+x), donc non identiquement nul.

Venons maintenant à la deuxième voie annoncée plus haut. Du fait de l'invariance par **G**, les blocs ($Q_h \times Q_{h'}$) de S ont une structure très simple, et la structure des sous–espaces invariants est claire. Si ($h \neq h'$), i.e. dans un bloc extradiagonal, tous les coefficients de la matrice sont égaux; si ($h=h'$), i.e. dans un bloc carré diagonal ($Q_h \times Q_h$), tous les termes diagonaux sont égaux entre eux ainsi que tous les termes extradiagonaux; et la valeur commune de ceux-là, diminuée de la valeur commune de ceux-ci, n'est autre que la valeur propre de S afférente au sous–espace E_h.

Pour les calculs numériques effectifs, il peut être avantageux de revenir au tableau de BURT.

Considérons d'abord les facteurs, associés à l'espace E_0. On les obtient en analysant le tableau (2p × 2p) obtenu à partir du tableau de BURT, (2n × 2n), en cumulant les p blocs de lignes et colonnes $Q+_h$ et $Q-_h$.

Pour les facteurs associés à un espace E_h ($h \neq 0$), il suffit de trouver à quelle valeur propre est relatif l'un de ces facteurs; qu'on peut choisir très simple. En effet, soit {q, q'} deux questions du niveau x_h: on cherche une fonction f sur J, dont le support est restreint aux quatre modalités {q+, q'+, q-, q'-} et qui est vecteur propre pour la transition associée au tableau de BURT. De ce qui précède, il résulte que l'on peut poser: f(q+) = - f(q'+); f(q-) = -f(q'-); et le rapport entre f(q+) et f(q-) est fixé par la condition générale que le barycentre des

modalités d'une question q, tombe à l'origine. On posera donc, en reprenant les notations de la fin du §2:

$$f(q+) = (1 - \mu_h) \ ; \ f(q'+) = - (1 - \mu_h) \ ; \ f(q-) = - \mu_h \ ; \ f(q'-) = \mu_h \ ;$$

(dans la suite, on écrira μ au lieu de μ_h); la valeur propre à laquelle ce facteur f (non normalisé) est associé dans l'analyse du tableau en (0,1), n'est autre que le rapport à f de son image par la transition associée au tableau de BURT; c'est-à-dire, e.g., le rapport à $f(q+)$ de la moyenne de la fonction f sur le profil de la ligne q+ du tableau de BURT.

Avec la normalisation adoptée au §2, *in fine*,la ligne q+ a pour total n.μ; la moyenne de f sur le profil de q+ est le quotient par n.μ de la somme:

$$(f(q+).k(q+,q+) + f(q-).k(q+,q-) + f(q'+).k(q+,q'+) + f(q'-).k(q+,q'-) \ ;$$

soit: $(1-\mu).\mu + 0 + (\mu-1).(3.\mu-1).(1/2) + \mu.(1-\mu).(1/2) = (1-\mu)/2$;

d'où, pour la valeur propre après division par $f(q+).n.\mu$: $(1 / (2.n.\mu))$.

Ce résultat offre matière à vérification: on peut refaire le calcul en considérant le rapport à $f(q-)$ de la moyenne de la fonction f sur le profil de la ligne q- du tableau de BURT. De plus, il est satisfaisant que la contribution à la trace des (n_h-1) facteurs associés à un sous-espace E_h, $(h\neq0)$, tende vers une limite finie $(n_h/2.n.\mu_h)$ quand le nombre des questions tend vers l'infini, mais que reste constante la répartition de leurs niveaux, fixée par les rapports (n_h/n), et la répartition des sujets, fixée par les poids m_h, qui déterminent les μ_h.

4 Retour à l'analyse de données réelles

Du point de vue de l'analyse des données les facteurs associés aux sous-espaces E_h, $(h\neq0)$ sont sans intérêt: il rendent compte seulement de la structure stratifiée du modèle; en revanche, les facteurs associés à E_0 ont l'intérêt de montrer quel est l'étagement propre au cas considéré.

Dans un cas réel, il ne faut pas s'attendre à ce que les données soient rigoureusement conformes au modèle stratifié; mais on peut conjecturer que, dans de nombreux cas, le modèle stratifié rendra compte d'un premier facteur de niveau général; et peut-être du plan (1, 2). Viendrait ensuite une typologie des questions par thèmes.