Les Cahiers de l'Analyse des Données
Vol. XIX - 1994 - n°3 - pp.335-342.

EFFETS DE LA REDONDANCE SUR L'ANALYSE D'UN QUESTIONNAIRE HÉTÉROGÈNE

[HÉTÉROGÈNE]

J. TCHOUANKAM (*)

1 Le questionnaire hétérogène, échantillon de plusieurs champs sémantiques

Le traité de "*L'Analyse des Données*" (TI.A, n°2, §1.3) formule, quant au tableau de base, deux exigences:

Homogénéité: toutes les grandeurs recensées dans le tableau sont des quantités de même nature.

Exhaustivité: les ensembles de marge I et J présentent un inventaire complet d'un domaine réel dont le cadre n'est guère discutable.

Bien satisfaites par un tableau de contingence tel que celui d'un recensement, ces exigences ne définissent, pour d'autres études, qu'un idéal plus ou moins inaccessible. En particulier, dans une enquête, il semble assez facile de délimiter la population visée, dont I sera un échantillon; mais l'ensemble Q des questions, restreint, *a priori*, par la capacité limitée des enquêtés et des enquêteurs, ne peut prétendre couvrir ni échantillonner un domaine qui n'a pas d'existence explicite: que comprendre, en effet, dans 'le mode de vie'; ou dans 'la connaissance des nouvelles technologies'? Bien plus, l'objet même de l'enquête peut être de confronter des données de plusieurs ordres; par exemple: condition sociale du sujet, connaissances objectives, opinions et suggestions.

En termes de sémantique, on parle de redondance; mathématiquement, il s'agit plutôt de pondérations: à supposer qu'une centaine de questions, empiétant plus ou moins entre elles, suffisent pour tout sonder, c'est-à-dire pour recouvrir un domaine continu d'informations, la trame du questionnaire est, sur ce domaine d'une inégale densité. L'on simulera donc la diversité des résultats d'analyse issus de divers questionnaires en analysant un seul tableau de base dont les colonnes reçoivent des pondérations variables; et l'ensemble des lignes (individus) est diversement composé.

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

Les cahiers de l'analyse des données 0339-3097/94/03/08/$ 4.00/© Gauthier-Villars

Notre propos n'est pas de démontrer l'adéquation à l'économie ou à la sociologie d'un tel schéma général; mais d'étudier mathématiquement quelques exemples; ce qui suffira à justifier des mises en garde que nous adresserons à ceux qui construisent et analysent des tableaux.

2 Modèle d'un questionnaire portant sur deux domaines indépendants

2.1 Ensembles de variables et formules de distance

Nous partons de deux ensembles de variables J_a et J_b; et supposons qu'un individu *i* est décrit, suivant J_a, par une ligne $\{k(i, j_a) \mid j_a \in J_a\}$ de total 1; ou, si l'on préfère, par un profil $f_{Ja}{}^{i}$ sur J_a, avec: $\forall j_a$: $f_{Ja}{}^{i} = k(i, j_a)$. Et de même, pour la description de i suivant J_b, un profil $f_{Jb}{}^{i}$. En termes concrets, on peut considérer J_a (ou J_b) comme l'ensemble des modalités de réponse à un ensemble Q_a de questions; (ou comme les modalités d'un ensemble Y_a de variables éclatées;) le total des composantes d'un bloc étant normalisé à 1, afin de simplifier les calculs et les notations.

Pour une population I_a, on a un tableau de correspondance $k(I_a, J_a)$; avec une loi marginale déterminée, f_{Ja}; d'où une métrique du χ^2 dans l'espace des profils sur J_a; et, pour le nuage $N(I_a)$, un système orthonormé d'axes principaux d'inertie; les cordonnées sur ces axes pouvant être notées $\{F_\alpha \mid \alpha \in A\}$; avec entre les profils de deux individus i_a et i'_a, une distance donnée par la formule:

$$d^2(i_a, i'_a) = \|f_{Ja}{}^{ia} - f_{Ja}{}^{i'a}\|^2 = \sum\{((F_\alpha(i_a) - F_\alpha(i'_a))^2 \mid \alpha \in A\} \ ;$$

la variance de F_α étant, comme d'usage, notée λ_α.

Des formules en tout analogues peuvent être écrites relativement à J_b, en remplaçant les lettres $\{a, \alpha, A\}$ par $\{b, \beta, B\}$.

2.2 Description d'un ensemble produit suivant deux ensembles de variables

Considérons maintenant un ensemble I dont chaque élément $i = (i_a, i_b)$ est une paire d'individus pouvant être décrits, respectivement, par les variables de J_a et de J_b. On aura un tableau de correspondance $I \times J = I \times (J_a \cup J_b)$.

Afin d'introduire des pondérations relatives entre J_a et J_b, nous posons:

$$\forall\, i = (i_a, i_b) : k(i, j_a) = t_a.k(i_a, j_a) \ ; \ k(i, j_b) = t_b.k(i_b, j_b) \ ;$$

où t_a et t_b sont des nombres réels positifs de somme 1.

Selon cette convention, le profil de i sur J, s'obtient en mettant bout à

bout les profils de i_a et i_b sur J_a et J_b, multipliés respectivement par t_a et t_b; ce qu'on écrira succintement:

$$f_J{}^i = t_a.f_{Ja}{}^{ia} + t_b.f_{Jb}{}^{ib} \ .$$

Supposons de plus que le profil de marge sur J soit, de même, donné par la formule:

$$f_J = t_a.f_{Ja} + t_b.f_{Jb} \ .$$

Alors entre $i=(i_a, i_b)$ et $i'=(i'_a, i'_b)$, on a la formule de distance:

$$d^2(i, i') = t_a.d^2(i_a, i'_a) + t_b.d^2(i_b, i'_b) \ ;$$

En effet, en bref, la formule de la distance du χ^2 sur J, comprend deux blocs de termes indicés, respectivement, par J_a et J_b. Considérons, v.g., le premier bloc: dans le profil de i sur J, les composantes sont celles du profil de i_a sur J_a, multipliées par t_a; de même pour le profil de marge. Ainsi, dans chaque terme de la distance du χ^2, le numérateur est multiplié par $(t_a)^2$ et le dénominateur par t_a; donc le terme est multiplié par t_a; CQFD.

En termes de facteurs pour i_a et i_b, la distance peut encore s'écrire:

$$d^2(i, i') = t_a.\Sigma\{((F_\alpha(i_a) - F_\alpha(i'_a))^2 \mid \alpha \in A\} + t_b.\Sigma\{((F_\beta(i_b) - F_\beta(i'_b))^2 \mid \beta \in B\} \ .$$

2.3 Analyse dans le cas de l'indépendance entre les deux ensembles de variables

Telle quelle, cette formule ne donne pas, en général, les résultats de l'analyse factorielle de la correspondance entre I et J; mais elle les donne sous l'hypothèse complémentaire que $I = I_a \times I_b$; ce qui équivaut à une hypothèse d'indépendance entre les deux ensembles de variables.

En effet, on a:

$$\forall \alpha \in A, \forall \beta \in B : \Sigma\{F_\alpha(i_a).F_\beta(i_b) \mid i_a \in I_a \ ; \ i_b \in I_b\} = 0 \ ;$$

la moyenne du produit étant égale au produit des moyennes.

Afin de retrouver les facteurs issus de la correspondance entre I et J, notons:

$$\forall i = (i_a, i_b) \in I, \forall \alpha \in A, \forall \beta \in B :$$
$$F_\alpha(i) = \sqrt{(t_a)}.F_\alpha(i_a) \ ; \ F_\beta(i) = \sqrt{(t_b)}.F_\beta(i_b) \ ;$$

Sur I, les fonctions $\{F_\alpha \ F_\beta\}$ sont de moyenne nulle et non corrélées entre elles, avec pour variances respectives $t_a.\lambda_\alpha$, $t_b.\lambda_\beta$. La formule de distance est:

$$d^2(i, i') = \Sigma\{(((F_\alpha(i) - F_\alpha(i'))^2 \mid \alpha \in A\} + \Sigma\{((F_\beta(i) - F_\beta(i'))^2 \mid \beta \in B\}.$$

Les facteurs des correspondances $I_a \times J_a$ et $I_b \times J_b$ se retrouvent donc; mais avec une autre échelle et d'autres valeurs propres; de façon précise, celles-ci sont multipliées respectivement par t_a et t_b (nombres dont on rappelle que la somme est 1).

Dans la suite des facteurs issus de $I \times J$, les facteurs des deux origines se mêlent; mais l'ordre d'ensemble dépend essentiellement des valeurs des pondérations t_a et t_b: si t_a tend vers 1 (donc t_b vers 0), les facteurs issus de $I_a \times J_a$ prennent les premiers rangs, devant ceux issus de $I_b \times J_b$; c'est le contraire si t_b tend vers 1.

Quant à des données réelles, l'interprétation du modèle proposé est claire: J_a et J_b décrivent deux aspects indépendants d'une même réalité; v.g., des opinions des mêmes sujets dans deux domaines qui n'ont entre eux aucun rapport. Chacun des facteurs obtenus concerne un seul des deux aspects. L'ordre dans lequel sortent les facteurs résulte de la pondération relative de J_a et J_b.

3 Questionnaire portant sur deux domaines distincts mais corrélés entre eux

3.1 L'antinomie entre indépendance et corrélation

Le titre de ce § est ambigu: comment concevoir une corrélation entre deux domaines distincts?

Soit l'ensemble I des individus, assimilé à un espace probabilisé (i.e., en bref, un ensemble muni d'une loi, ou mesure positive de masse totale 1; dans le cas le plus simple, chaque individu i d'un ensemble fini I reçoit la masse 1/cardI): une variable aléatoire est un vecteur de l'espace $L^2(I)$ des fonctions de carré sommable sur l'espace probabilisé de base.

On se représentera alors un domaine D_a, de la réalité afférente à I, comme un ensemble de variables; indéfini, en ce que les mesures ou questions peuvent être renouvelées indéfiniment; mais caractérisé par le fait que les vecteurs représentant ces variables sont dans un sous-espace fixé, L_a, de $L^2(I)$. À un autre domaine, D_b, sera associé un autre sous-espace, L_b.

On dira que les domaines D_a et D_b sont corrélés entre eux si, au sein de $L^2(I)$, les deux sous-espaces, L_a et L_b, ne sont pas orthogonaux entre eux. En termes de statistique, l'étude simultanée de D_a et D_b relève de l'analyse canonique; en géométrie, on recourra à l'étude, bien connue, de la figure formée par deux sous-espaces d'un espace euclidien.

L'analyse extrait les dimensions communes à D_a et D_b; c'est-à-dire les couples de vecteurs de L_a et L_b formant un angle dont le cosinus est maximum. La signification propre à chacun des deux domaines, n'intervient

pas directement, non plus que leur importance relative: c'est pourquoi, dans ce modèle, la redondance, dont le §2 nous a montré le rôle essentiel, semble sans effet.

Dans l'esprit des psychologues et des sociologues qui, notamment depuis l'avènement des ordinateurs, tentent de fonder sur des modèles mathématiques la conception des questionnaires, sont présentes deux exigences contradictoires: d'une part, mesurer comme des entités indépendantes des dimensions cachées dont le langage commun suggère l'existence: intelligence, activité, docilité; d'autre part, calculer, entre ces entités, des corrélations qui, sous la condition d'indépendance, ne peuvent qu'être nulles.

L'analyse des correspondances ne résoud cette antinomie que dans la mesure où les domaines D_a et D_b (e.g. comportement verbal et qualités somatiques) ont véritablement une intersection vide: car, autrement, les premiers facteurs seront trivialement créés par les items communs aux deux blocs mis en correspondance; ce qu'on peut appeler un effet de redondance inter-bloc.

3.2 Un cas modèle avec deux ensembles isomorphes de variables

À titre d'exercice, on propose ci-après un modèle, issu de celui du §2, où s'introduit une certaine corrélation entre deux blocs, supposés isomorphes.

Les hypothèses des §§2.1 et 2.2 sont conservées telles quelles; et complétées comme suit.

A) Les ensembles $I_a \times J_a$ et $I_b \times J_b$ se correspondent biunivoquement avec les valeurs des deux tableaux $k(I_a, J_a)$ et $k(I_b, J_b)$; le modèle commun pouvant être noté $k(I_c, J_c)$. Cette hypothèse pourrait être réalisée dans le cas de deux échelles de GUTTMAN, a et b, de même structure. L'ensemble des facteurs issus de $k(I_c, J_c)$ sera indicé par $\gamma \in \Gamma \approx A \approx B$; et les valeurs propres seront notées λ_γ.

B) L'ensemble I est réunion de deux sous-ensembles, I_p et I_d:

$I_p = I_a \times I_b \approx I_c \times I_c$, est un produit, comme l'ensemble I du §2.3 ;

$I_d = \{(i_c, i_c) \mid i_c \in I\}$, constitue la diagonale de ce produit.

C) Les pondérations relatives des ensembles I_p et I_d sont définies par des coefficients r_p et r_d, de somme 1. Pour la simplicité des calculs, on supposera d'abord que les deux ensembles de variables reçoivent même poids:

$t_a = t_b = (1/2)$;

le cas général sera traité ensuite (cf. *infra*, §3.3).

Sous ces hypothèses, comme dans la structure particulière considérée au §2.3 (où r_d=0), le profil de marge sur J est donné par la formule:

$$f_J = t_a.f_{Ja} + t_b.f_{Jb} \ ;$$

la distance du χ^2 entre éléments de I subsiste donc:

$$d^2(i, i') = t_a.\Sigma\{((F_\alpha(i_a) - F_\alpha(i'_a))^2 \mid \alpha \in \Gamma\} + t_b.\Sigma\{((F_\beta(i_b) - F_\beta(i'_b))^2 \mid \beta \in \Gamma\} \ ;$$

où $t_a = t_b = (1/2)$.

Afin de retrouver les facteurs issus de la correspondance entre I et J, notons:

$\forall\, i = (i_a, i_b) \in I, \forall\, \gamma \in \Gamma$:

$F_{\gamma+}(i) = (1/2).((F_\gamma(i_a) + F_\gamma(i_b))$;

$F_{\gamma-}(i) = (1/2).((F_\gamma(i_a) - F_\gamma(i_b))$;

On ainsi un système de coordonnées orthonormé; car la distance s'écrit:

$$d^2(i, i') = \Sigma\{(F_{\gamma+}(i) - F_{\gamma+}(i'))^2 + (F_{\gamma-}(i) - F_{\gamma-}(i'))^2 \mid \gamma \in \Gamma\}.$$

Il est clair que, pour $\gamma \neq \gamma'$, le produit de deux coordonnées $(F_{\gamma\pm}, F_{\gamma\pm})$ a une moyenne nulle quels que soient les signes; reste donc le cas $\gamma=\gamma'$.

Sur le produit I_p, on a:

$$moy\{F_{\gamma+}.F_{\gamma+} \mid i \in I_p\} = moy\{F_{\gamma-}.F_{\gamma-}\} = \lambda_\gamma/2 \ ; \ moy\{F_{\gamma+}.F_{\gamma-}\} = 0 \ ;$$

sur la diagonale I_d, compte tenu de ce que $i_a=i_b$, d'où $F_{\gamma-}=0$, on a:

$$moy\{F_{\gamma+}.F_{\gamma+} \mid i \in I_d\} = \lambda_\gamma \ ; \ moy\{F_{\gamma-}.F_{\gamma-}\} = moy\{F_{\gamma+}.F_{\gamma-}\} = 0 \ ;$$

d'où sur I :

$$moy\{F_{\gamma+}.F_{\gamma+} \mid i \in I\} = (r_d+(r_p/2)).l_\gamma \ ;$$

$$moy\{F_{\gamma-}.F_{\gamma-} \mid i \in I\} = (r_p/2).l_\gamma \quad ; \quad moy\{F_{\gamma+}.F_{\gamma-}\} = 0 \ ;$$

On a donc deux groupes de facteurs dont chacun est indicé par γ: les facteurs $F_{\gamma+}$ et $F_{\gamma-}$, afférents aux valeurs propres $(r_d+(r_p/2)).\lambda_\gamma$ et $(r_p/2).\lambda_\gamma$; ces facteurs seront dits respectivement directs: $F_+((i_c, i'_c)) = F_+((i'_c, i_c))$; et inverses: $F_-((i_c, i'_c)) = - F_-((i'_c, i_c))$. Si r_p est nul, les facteurs directs disparaissent; si r_d est nul, (comme au §2.3,) facteurs direts et inverses vont par paires dans la suite des valeurs propres; en général, les deux groupes de facteurs peuvent se mêler de façon quelconque.

3.3 NOTE: Résolution du modèle dans le cas général

Dans le cas général $t_a \neq t_b$, le calcul se fait comme suit. On note:

$$t_a = (\cos\vartheta)^2 \ ; \ t_b = (\sin\vartheta)^2 \ ;$$

au lieu de $F_{\gamma+}$ et $F_{\gamma-}$, on introduit, pour coordonnées, des combinaison linéaires quelconques:

$$G_\gamma(i) = (u.F_\gamma(i_a) + v.F_\gamma(i_b)) \ ; \ H_\gamma(i) = (u'.F_\gamma(i_a) + v'.F_\gamma(i_b)) \ ;$$

pour que les coordonnées $G_\gamma(i)$ et $H_\gamma(i)$ soient orthonormées, il faut que:

$$G_\gamma^2 + H_\gamma^2 = t_a.F_\gamma(i_a)^2 + t_b.F_\gamma(i_b)^2 \ ;$$

soit, relativement à u, v, u', v' :

$$u^2 + u'^2 = t_a \ ; \ v^2 + v'^2 = t_b \ ;$$

d'où le paramétrage trigonométrique:

$$G_\gamma(i) = \cos\varphi.\cos\vartheta.F_\gamma(i_a) + \sin\varphi.\sin\vartheta.F_\gamma(i_b) \ ;$$

$$H_\gamma(i) = \sin\varphi.\cos\vartheta.F_\gamma(i_a) - \cos\varphi.\sin\vartheta.F_\gamma(i_b) \ ;$$

l'angle φ est fixé en demandant un comoment nul de $G_\gamma.H_\gamma$ sur I; soit, en éliminant d'emblée le facteur λ_γ:

$$0 = r_p.(\sin\varphi.\cos\varphi.\cos^2\vartheta - \sin\varphi.\cos\varphi.\sin^2\vartheta) +$$
$$r_d.(\cos\varphi.\cos\vartheta + \sin\varphi.\sin\vartheta).(\sin\varphi.\cos\vartheta - \cos\varphi.\sin\vartheta) \ ;$$

d'où, en multipliant par 2 :

$$0 = r_p.(\sin(2\varphi) \ . \ \cos(2\vartheta)) \ +$$
$$r_d.((\sin(2\varphi) \ . \ \cos(2\vartheta)) - (\cos(2\varphi) \ . \ \sin(2\vartheta))) \ ;$$

d'où :

$$0 = (r_p+r_d).(\sin(2\varphi) \ . \ \cos(2\vartheta)) \ - \ r_d.(\cos(2\varphi) \ . \ \sin(2\vartheta)) \ ;$$

d'où :

$$\mathrm{tg}(2\varphi) = (r_d/(r_p+r_d)).\mathrm{tg}(2\vartheta) \ ;$$

le cas particulier traité d'abord est $\vartheta = \pi/4$; $2.\vartheta = 2.\varphi = \pi/2$.

Les valeurs propres afférentes à G_γ et H_γ se calculent aisément comme les moments des carrés de ces combinaisons des $F_\gamma(i_a)$, $F_\gamma(i_b)$.

4 Sommaire et conclusion

On peut regarder les variantes de l'analyse d'une correspondance $I \times J$, comme différant entre elles par le choix aléatoire d'individus et de variables, extraits d'espaces potentiels déterminés; ou encore, par l'introduction de lois de pondération arbitraires sur ces espaces eux-mêmes, mis en correspondance.

Tandis que le choix des individus se fait, généralement, de façon explicite, il n'en est pas de même pour les variables; auxquelles on demande seulement d'avoir rapport à tous les aspects du problème étudié.

En particulier, dans le cas d'un ensemble J hétérogène, on peut supposer que la pondération au sein de sous-groupes homogènes J_a, J_b,... est mieux maîtrisée que la pondération relative entre les sous-groupes.

Selon cette hypothèse, ayant fixé deux blocs J_a et J_b, on considère des cas modèle: l'un (§2) où il y a indépendance entre J_a et J_b; l'autre (§3) où le choix de I introduit une corrélation entre les blocs. Il apparaît sur ces modèles que l'ordre dans lequel l'information est extraite, de $I \times J$, sous forme de facteurs, dépend essentiellement des pondérations, de la redondance.

L'étude de la correspondance entre deux blocs, J_a et J_b, de variables est en butte à des effet plus facheux encore: si, explicitement ou implicitement, les deux blocs ont une intersection non vide (e.g., dans une enquête, du fait que l'âge des sujets est omniprésent), l'analyse est dominée par cette redondance inter-bloc, et ne montre aucunement les corrélations non triviales entre J_a et J_b.

D'ordinaire, le statisticien reçoit des données déjà recueillies qu'il doit analyser: mais il reste libre de maîtriser les effets de la redondance par un choix approprié des blocs de variables et l'introduction éventuelle de pondérations *a posteriori*.