LES CAHIERS DE L'ANALYSE DES DONNÉES

D. MAÏTI

Effet Guttman sur le tableau de Burt associé à des variables fortement liées, modèle général et exemple d'application

Les cahiers de l'analyse des données, tome 4, n° 3 (1979), p. 261-270

<http://www.numdam.org/item?id=CAD_1979__4_3_261_0>

Les Cahiers de l'Analyse des Données
Vol. IV - 1979 - n° 3 - p. 261-270

EFFET GUTTMAN SUR LE TABLEAU DE BURT ASSOCIÉ À DES VARIABLES FORTEMENT LIÉES, MODÈLE GÉNÉRAL ET EXEMPLE D'APPLICATION [BURT GUTTMAN]

par D. Maïti (1)

L'analyse du tableau de Burt associé à des variables fortement liées, produit d'abord plusieurs facteurs rendant compte de cette liaison ; et au-delà, la structure devient confuse, ne se prêtant à aucune interprétation. Pourtant dans le cas de données réelles, la liaison est rarement si forte qu'elle ne laisse subsister outre une forte structure unidimensionnelle, des variations transversales intéressantes. Dans cette note, on étudie d'abord l'analyse du tableau de Burt dans un cas modèle purement unidimensionnel (§1.1), et dans un cas réel qui est à l'origine du présent travail (§1.2). Puis (§2) on propose d'effectuer sur les données plusieurs transformations, afin d'en extraire toute la structure (§§ 2.1, 2.2 et 2.4) et on rend compte d'un essai portant sur des données réelles (§ 2.3). Au passage (§1.3), à partir des bornes des classes d'un codage disjonctif complet, on associe une courte représentative précise, à une relation approchée entre deux ou plusieurs variables.

1 *Analyse du tableau de Burt associé à des variables fortement liées* :

1.1 Etude d'un cas modèle : reprenons les notations usuelles du codage sous forme disjonctive complète pour un ensemble I d'individus décrit par un ensemble Q de p variables numériques ; nous supposons que l'intervalle de variation d'une variable q est partagé en un ensemble J_q de sous intervalles consécutifs dont chacun contient les coordonnées de card I/n individus (e.g. si n = 4, card I = 100 ; les bornes des sous- intervalles étant 0,7 ; 0,9 et 1,15 : il y a 25 individus i pour lesquels la coordonnée x_q (i) est inférieure à 0,7 ; 25 pour lesquels elle est comprise entre 0,7 et 0,9 ; etc...)
L'ensemble J_q est encore appelé : ensemble des modalités de la variable q ; et on dit qu'un individu i possède la modalité $j \in J_q$, si sa coordonnée $x_q(i)$ tombe dans l'intervalle j, on note :

$$J = \cup\{J_q \mid q \in Q\} \; ; \; \text{Card}\, J = p \times n$$

$k(i,j) = 1$ si l'individu i possède la modalité j ; et zéro sinon

$b(j,j') = \Sigma\{k(i,j)\, k(i,j') \mid i \in I\}$ = nombre des individus de I, possédant à la fois les modalités j et j'.

Le tableau k_{IJ} est appelé tableau de description sous forme disjonctive complète ; et le tableau b_{JJ}, tableau de Burt associé à celui-ci. L'analyse des deux tableaux fournit sur J les mêmes facteurs φ^J de variance 1, mais avec des valeurs propres différentes : λ pour k_{IJ} et $\Lambda = \lambda^2$ pour b_{JJ}. Si on note f^J_J la transition associée au tableau de Burt l'équation des facteurs s'écrit : $\varphi^J \circ f^J_J = \lambda\, \varphi^J$. (pour plus de précision on

(1) Laboratoire de physique corpusculaire. Collège de France.
Laboratoire de statistique. Université P. et M. Curie.

se reportera à l'article (BIN. MULT), *cahiers* vol. II n° 1 pp. 55 sqq, 1977 ; ou à la leçon VI n° 0 du livre ENS 2, Dunod ed. 1979).

Ceci étant rappelé, nous disons que les variables de Q sont fortement liées entre elles, si chacun des blocs $J_q \times J_{q'}$ du tableau de Burt est fortement concentré autour de sa diagonale : dans ce cas en effet, la plupart des individus pour lesquels e. g. la variable q tombe dans le 3° sous-intervalle, ont aussi la variable q'dans le 3° sous-intervalle ou dans des intervalles adjacents : le 2° ou le 4° principalement. Il faut prendre garde que le 3° sous-intervalle de la variable q, n'a pas les mêmes bornes que le 3° intervalle de la variable q' : il s'agit simplement d'une relation fonctionnelle approchée entre x_q et $x_{q'}$; non d'une identité.

Figure 1 : cas modèles de 3 variables fortement liées entre elles

Sur la figure, on a schématisé le tableau de Burt dans le cas de 3 variables (p = 3) fortement liées entre elles. Les blocs diagonaux, e.g. $J_{q_1} \times J_{q_1}$ du tableau de Burt, sont comme de règle , réduits à leur diagonale (car si j, j' $\in J_q$ et j $\neq$ j', il est impossible qu'un même individu possède à la fois les modalités j et j'. Quant aux autres blocs, $J_q \times J_{q'}$ (avec q $\neq$ q'), on les a figurés par des bandes diagonales hachurées. Pour résoudre complètement l'analyse nous supposerons que les tableaux $J_q \times J_{q'}$ (q $\neq$ q') sont des tableaux carrés symétriques tous égaux entre eux.

Dans ces conditions, notons φ^{Jq} un facteur issu d'un tel tableau carré et satisfaisant à l'équation : $\varphi^{Jq} \circ t^{Jq'}_{Jq} = \lambda' \varphi^{Jq'}$; où on a noté t la transition associée à un bloc carré extra diagonal de b_{JJ}. Il est facile de vérifier que la fonction $\varphi^J = \{1,1,1\} \otimes \varphi^{Jq}$, réalisée en mettant bout à bout p exemplaires de φ^{Jq} (3 dans le cas de la figure) est un facteur relatif à la valeur propre $\lambda' = ((p-1)\lambda + 1)/p$: en effet, en bref la matrice $p f^J_J$ (matrice de la transition associée à b_{JJ}, multipliée par p), comprend d'une part des blocs diagonaux $J_q \times J_q$ qui sont des matrices identité, et d'autre part des blocs extradiagonaux égaux à $t^{Jq'}_{Jq}$. Si maintenant on met bout à bout p exemplaires de φ^{Jq}, multipliés par des coefficients a_q de moyenne nulle, on obtient une fonction $\varphi^J = \{a_1, a_2, a_3\} \otimes \varphi^{Jq}$, qui est un facteur relatif à la v.p. $\lambda'' = (1-\lambda)/p$. On a ainsi tous les facteurs issus de b_{JJ} (ou de k_{IJ}). Chacun des facteurs φ^{Jq} issus de $t^{Jq'}_{Jq}$ et relatifs à λ, donne un facteur relatif à $\lambda' = ((p-1)\lambda + 1)/p$, et (p - 1) facteurs relatifs

à $\lambda'' = (1-\lambda)/p$. Sur l'histogramme des valeurs propres, on a d'abord les n valeurs propres λ' (n = card J_q) qui s'échelonnent depuis 1 (facteur trivial) jusqu'à (1/p) (qui correspond à $\lambda = 0$) ; puis les v. p. multiples λ'', qui vont de (1/p) (qui correspond à $\lambda = 0$) jusqu'à 0 (v.p. dont la multiplicité est p-1, dans toute analyse de tableau de Burt avec Card Q = p). Sur la figure 1, p = 3. Pour tracer l'histogramme des v.p. associées à l'analyse du tableau k_{IJ}, on a posé n = 5, et fixé arbitrairement pour valeurs de λ : 1 ; 0,85 ; 0,7 ; 0,55 ; 0,25. D'où pour les λ' : 1 ; 0,9 ; 0,8 ; 0,7 ; 0,5 ; et pour les λ'' : 0 ; 0,05 ; 0,1 ; 0,15 ; 0,25. La coupure étant à (1/p) ≈ 0,33.

1.2 *Exemple de données concrètes* : dans l'étude statistique d'une expérience de physique corpusculaire à très haute énergie (cf. [HODOGRAPHES] *in Cah.* Vol. IV n° 3) on a considéré 259 événements ayant produit au total 3508 particules enregistrées. Parmi celles-ci, 3387 sont considérées comme des sommets de l'hodographe des événements qui les a produits et on a calculé pour ces sommets cinq variables numériques notées DA, DB, SL, SS, DT. (cf. [ANGL. SOMM. CONV.], même cahier). Le tableau de Burt, construit en divisant l'intervalle de variation de chaque variable en 10 sous-intervalles d'égale fréquence (cf. *supra* §1.1), est conforme au schéma de la figure 2 ; schéma qui ne diffère de la figure 1 qu'en ce que dans certains blocs $J_q \times J_{q'}$, les cases les plus chargées s'alignent sur la diagonale ascendante parce que les variables q et q' sont fortement liées, mais varient en sens opposé.

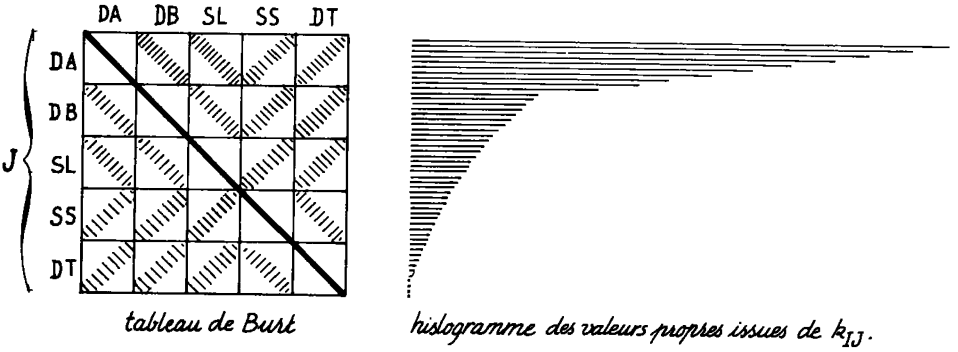


Figure 2 : cas réel de 5 variables fortement corrélées entre elles

Cela évidemment, ne perturbe en rien le modèle : il suffirait de changer le signe des variables SS et DT, ou de numéroter dans l'ordre inverse leurs modalités. Mais il faut noter que les bandes diagonales des blocs $J_q \times J_{q'}$ du tableau de Burt ne sont pas rigoureusement égales entre elles. Cependant, les résultats de l'analyse sont conformes au modèle du §1.1.

L'histogramme des v.p. montre d'abord une suite de 10 v.p. (la v.p. 1 y comprise) décroissant régulièrement jusqu'à 0,35, avec des intervalles compris entre 0,04 et 0,09, puis vient une brusque décroissance de 0,11 ; et les v.p. décroîssent alors de 0,24 à 0 avec des intervalles tous inférieurs à 0,022. La rupture est donc nette au niveau de la 10° v.p.. Quant aux facteurs, on a dans le plan 1 x 2 dix îlots dessinant une parabole (cf. fig. 3) ; et ces dix îlots sont : {DA1 ; DB1 ; SL1 ; SS10 ; DT10}, {DA2 ; DB2 ; SL2 ; SS9 ; DT9}, ... ; {DA10 ; DB10 ; SL10 ; SS1 ; DT1}. Ces mêmes îlots se retrouvent dans le plan 1 x 3, décrivant une cubique : ce net effet Guttman est également conforme au modèle pour lequel les dix premiers facteurs sont identiques à ceux issus d'un bloc 10 x 10 ($J_q \times J_{q'}$). Le nuage des

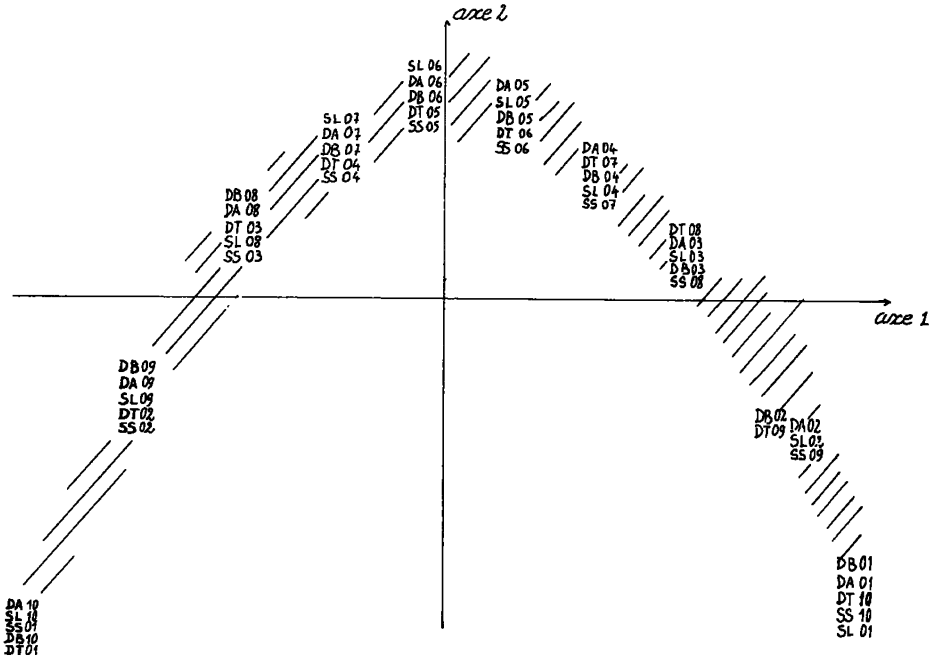
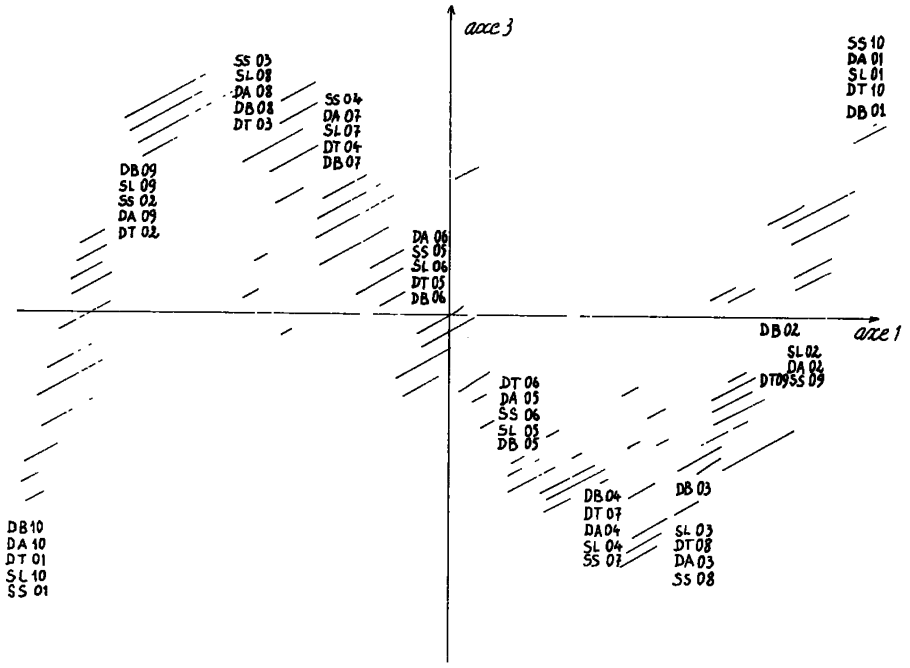


Figure 3 : plans 1x2 et 1x3 issus de l'analyse factorielle du tableau k_{IJ}.



individus (les 3387 sommets, que nous avons suggérés par des hachures), s'écarte peu lui-même de la courbe des modalités des variables.

1.3 *Relation fonctionnelle approchée entre variables fortement liées* :

Comme on l'explique dans la note [HODOGRAPHE], la forte liaison entre les cinq variables{DA, DB, SL, SS, DT}peut être justifiée par un modèle physique ; et ce modèle suggère de substituer aux variables primitives de nouvelles variables qui en sont des fonctions monotones : on a ainsi cinq angles {AA, BB, LL, UU, TT}qui tous sont compris entre 0 et $\pi/2$, et varient ensemble dans le même sens. Evidemment, il ne s'agit pas d'une relation fonctionnelle rigoureuse, vu l'aspect statistique du phénomène naturel lui-même ; mais il est intéressant de schématiser cette relation par une fonction.
Prenons l'exemple de la relation entre AA et BB. Le bloc correspondant du tableau de Burt est transcrit sur la figure 4. Dans cette figure, on n'a pas comme il est d'usage, donné aux lignes et colonnes une même largeur, mais on a disposé les cases du tableau dans le plan rapporté à deux axes AA et BB, et conservant les proportions des intervalles AA_i, BB_j ; par exemple, les bornes de la classe AA06 sont (0,431 ; 0,551) et celles de la classe BB07 sont (0,236 ; 0,324) ; on a donc inscrit dans

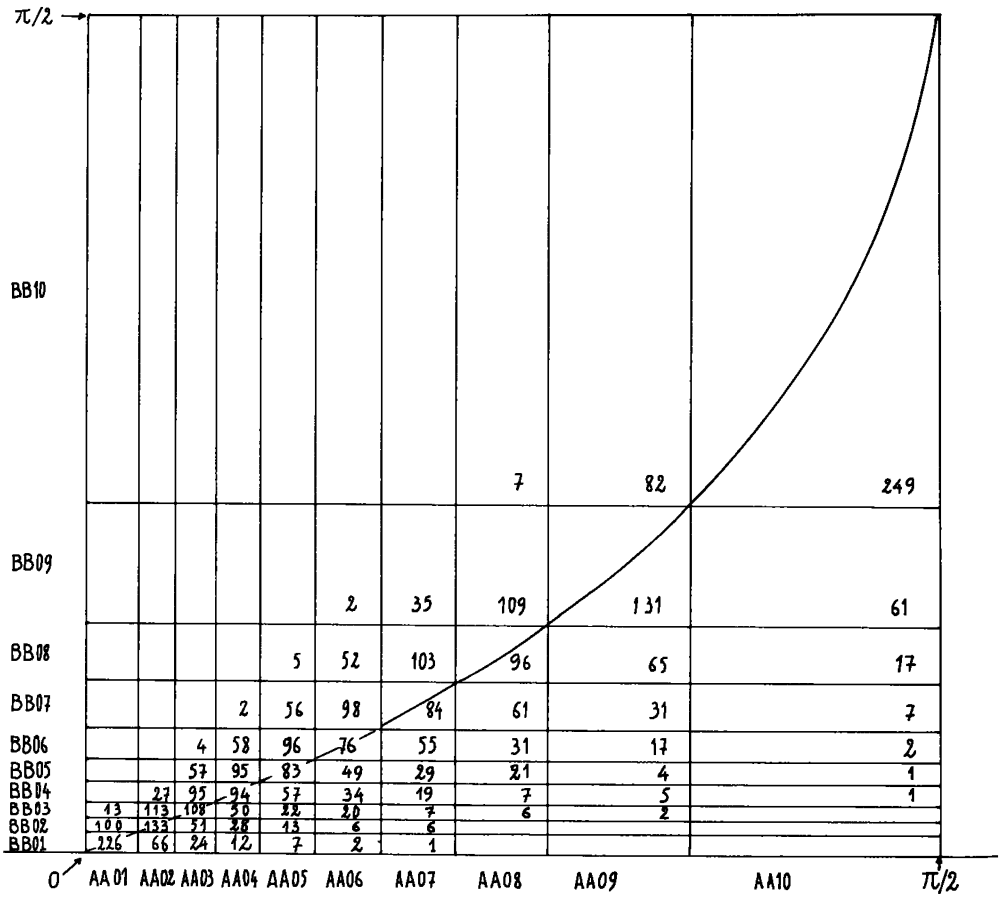


Figure 4 : Bloc AA × BB du tableau de Burt et courbe moyenne.

le rectangle produit de ces deux intervalles le contenu de la case k(AA06, BB07) du tableau de Burt : soit 98. Sans qu'il soit besoin de porter les 3387 sommets représentés par les points de coordonnées (AA, BB) (ceci serait toutefois possible) on a sur la figure 4 une image sensible de la densité du nuage de ces points. On sait de plus (cf. [ANGL. SOMM. CONV.]) dans le cas présent, que par définition, la coordonnée AA est supérieure à la coordonnée BB ; donc le nuage est tout entier au-dessous de la droite AA = BB.

Reste à expliquer la courbe de la figure 4. Supposons le bloc AA x BB du tableau de Burt, réduit à sa diagonale : sur la fig. 4, tous les points du nuage tomberaient dans la suite des cases (AA01 x BB01), (AA02 x BB02), etc... Les points de contact de ces cases ne sont autres que les points ayant pour abscisses et ordonnées les bornes des classes de même rang en AA et BB : e. g. le point (0,551 ; 0,236) dont l'abscisse et l'ordonnée sont respectivement (cf. *supra*) les bornes supérieures des classes AA06 et BB06. Ces points suggèrent une courbe ; les premiers points à partir de l'origine sont sensiblement alignés ; l'alignement est encore plus net si l'on prend pour coordonnées non les angles AA et BB, mais leurs tangentes. Il est facile de préciser cette courbe : en effet e.g. l'abscisse 0,551 n'est autre que la valeur de l'angle AA en dessous de laquelle on trouve les abscisses des (6 x 339) sommets des classes AA01 à AA06 ; et de même pour l'ordonnée 0,236 relativement aux classes BB01 à BB06. Plus généralement on peut noter AA(r) la valeur qui a le rang r dans l'ensemble des valeurs prises par l'angle AA; et de même pour BB(r). Nous avons tracé sur la figure 4 la courbe des points (AA(r) ; BB(r)) ; et avons également construit (par l'ordinateur évidemment) les courbes analogues relatives à tous les couples de variables. Ces courbes se ressemblent ; avec une partie linéaire encore plus nette si l'on utilise pour coordonnées les tangentes des angles.

Remarque : Comparaison entre ensembles de données réelles ou simulées : la présentation donnée ici des sous-tableaux du tableau de Burt, avec une courbe moyenne, permet de comparer les relations fonctionnelles approchées entre variables, présentes dans plusieurs ensembles de données analogues expérimentales ou simulées. Ces comparaisons sont assez fines pour mettre à l'épreuve des modèles théoriques jusqu'ici acceptés sans réticence ; et aussi pour suggérer des hypothèses physiques : en montrant, e.g. que deux classes de sommets caractérisées globalement par des intervalles de valeurs pour deux paramètres tels que l'énergie et le moment transverse, suivent ou non la même loi moyenne ; et si elles diffèrent, en quoi elles le font.

Cependant les comparaisons suggérées ici ne portent que sur la loi de la forme moyenne qui dépend d'un seul paramètre le rang : l'objet du § 2 est de décrire par l'analyse statistique la diversité des formes de même rang (i.e. ici : au voisinage de AA = 0,551 ; et BB = 0,236, on trouve des points pour lesquels le rapport AA/BB dépasse la valeur typique (0,551/0,236), ou lui est inférieure). Les comparaisons fondées sur ces nouveaux paramètres seront encore plus fines.

2. *Recherche de facteurs indépendants du rang* :

2.1 *Construction de données fictives en fonction du rang moyen* :

Dans ce §, de préférence aux notations mathématiques usuelles, nous prenons des notations proches de celles du langage FORTRAN. On note :

DIV (I, JV) : un tableau de données NI x NV décrivant un ensemble de NI individus par un ensemble de NV variables : nous supposons que ces variables sont fortement liées entre elles et varient dans le même sens.

NRV (I, JV) : le numéro (ou rang) de la quantité DIV (I, JV) (valeur de la variable JV, pour l'individu I) au sein de l'ensemble {DIV(IX, JV) | IX = 1,..., NI} , des valeurs prises par cette variable pour tous les individus. Pour JV fixé NRV (I, JV) prend une fois et une seule fois toutes les valeurs entières de 1 à NI ; ce qui permet

de définir ci-dessous la fonction NIV ;

NIV (IR, JV) = le nom (ou indice) de l'individu I pour lequel on a NRV (I, JV) = IR ; i.e. de l'individu I pour lequel la valeur de la variable JV a le rang IR. Pour JV fixé, NIV et NRV sont des fonctions inverses l'une de l'autre :

$$\forall\ I = 1,\ldots,\ NI\ :\ NIV(NRV(I,JV),\ JV) = I\ ;$$

$$\forall\ IR = 1,\ldots,\ NI\ :\ NRV(NIV(IR,JV),JV) = IR\ ;$$

en effet il y a correspondance biunivoque entre les numéros d'individus et les rangs.

$$MR(I) = Ent(\ \Sigma\{NRV(I,JV)\ |\ JV = 1,\ldots,\ NV\}/NV)\ ;$$

le nombre MR (I) est la partie entière de la moyenne des rangs NRV(I,JV) des valeurs des NV variables pour l'individu I.

FIV(I, JV) = DIV(NIV(MR(I),JV),JV) : FIV(I,JV) est appelé valeur fictive de la variable JV pour l'individu I : cette valeur est celle qui dans $\{DIV(IX,JV)|IX = 1,\ldots,\ NI\}$, (ensemble de toutes les valeurs prises par la variable JV) a pour rang le rang moyen MR (I). Si on effectue sur le tableau fictif FIV, la construction de la figure 4 (§ 1.3), tous les points tombent exactement sur ce qu'on a appelé la courbe moyenne. En effet pour les individus fictifs, toutes les variables JV sont des fonctions monotones d'une variable intermédiaire unique : le rang moyen.

NB : pour construire dans un temps acceptable le tableau NRV, il convient de disposer de programmes de tri rapide : nous nous réservons de revenir dans un autre article sur le principe de ces programmes très utiles en analyse des données.

2.2 *Comparaison des individus donnés aux individus fictifs* : une analyse telle que celle du § 1.2 portant sur des variables numériques fortement liées entre elles, montre un paramètre de forme unique, qui par un fort effet Guttman, domine tous les facteurs. Dans le cas de l'exemple, cf [HODOGRAPHE] §6, ce paramètre a été appelé acuité : parce qu'il oppose les faibles valeurs des angles (angles aigus, pointes du convexe) aux fortes valeurs. En tout cas, ce paramètre de forme est lié au rang moyen : dans l'espace des trois premiers axes, les points représentatifs des individus s'écartent peu d'une courbe moyenne ; où ils s'ordonnent à peu près suivant ce rang. Pour extraire de nouveaux facteurs de formes indépendants du rang moyen, on décrira chaque individu donné par son rapport à l'individu fictif qui lui a été associé. De façon précise, on pose :

$$QIV(I,JV) = DIV(I,JV)/FIV(I,JV)\ ;$$

(ce codage est inspiré par la relation quasilinéaire que montre la figure 4 : mais cf. § 2.4 , ce nest pas le seul possible). Ainsi un individu I présentera une valeur QIV(I,JV) supérieure à 1 si parmi les individus dont le rang moyen est voisin du sien, I se distingue par une valeur élevée de la variable donnée DIV (I, JV). Il importe de noter qu'une valeur faible ou élevée de QIV peut exister aussi bien pour un individu de rang moyen faible qu'un individu de rang moyen élevé : toutefois il n'est pas assuré que la dispersion des valeurs de QIV (I, JV) soit la même tout au long de l'échelle du rang moyen. Voyons ce qu'il en est dans le cas de l'exemple (§2.3), avant de suggérer des perfectionnements au codage (§2.4).

2.3 *Analyse comparative sur des données concrètes* : on part d'un tableau de données DIV avec NI = 3387, NV = 5 ; on construit le tableau FIV (tableau fictif, calculé sur les rangs moyens) et le tableau QIV des quotients comparatifs. Les 5 variables numériques
{QIV(.,JV) | JV = 1,..., 5} sont alors codées chacune en 10 modalités d'égal effectif. Ces modalités sont notées de QA01 à QA10 pour la première variable ; ... ; QT01 à QT10 pour la cinquième variable. Le rang moyen MR(I) est également codé en dix modalités notées de MR01 à MR10.

On a ainsi un tableau de description logique k_{IJ} à 3387 lignes et 60 colonnes. On analyse ce tableau en mettant en éléments supplémentaires les dix colonnes MR01 à MR10.
Considérons d'abord les variables principales : dans les plans 1 x 2 et 1 x 3, on a un net effet Guttman : les modalités des variables QA, QU, QT sont associées et régulièrement disposées de 01 à 10 ; les modalités des variables QB et QL vont ensemble, et se disposent dans le sens opposé à celui des modalités de QA, QU, QT. Toutefois le point QU01 (et à un moindre degré QA01) s'écarte de la ligne régulière dessinée par les autres modalités. Pour un essai d'interprétation géométrique et physique de l'opposition {QA ; QU ; QT} ≠ {QB ; QL}, nous renvoyons à[HODOGRAPHE] §7.

Quant au nuage des individus il est fort dispersé du côté positif du premier facteur ($F_1 > 0$) (faibles modalités de QA, QU, QT ; fortes modalités de QB, QL) ; mais du côté $F_1 < 0$, il suit de plus près la courbe décrite par les modalités des variables Q.

Dans le plan 1 x 2 (et aussi dans le plan 1 x 3), les modalités supplémentaires s'ordonnent régulièrement de MR01 à MR10. Les 9 premières modalités, (MR01 ; MR02 ;...) occupent une position centrale ; mais MR10 (et à un moindre degré MR09) est fortement attiré par les valeurs 4 à 7 des modalités Q. L'interprétation de ce fait est facile : pour les rangs moyens élevés, les variables angulaires réelles et les variables angulaires fictives sont toutes voisines de leur maximum $\pi/2$: leur rapport ne peut donc s'écarter de la valeur 1 qui correspond aux modalités moyennes.

2.4 *Proposition de nouvelles formes de codage* :

Dans la présente note on s'intéresse surtout au problème général de l'analyse des données fortement liées : de ce point de vue la position des points MR10 (et MR09) n'est pas pleinement satisfaisante : car elle témoigne que le nouveau facteur de forme extrait par la présente analyse n'est pas totalement indépendant du rang moyen : on désirerait au contraire voir tous les points MR resserrés autour de l'origine.

Plusieurs modifications sont possibles ; dans le cas des données de l'exemple, il serait peut-être préférable de travailler sur le tableau des tangentes plutôt que sur le tableau des angles : les tangentes en effet s'étendent de 0 à ∞, avec pour leurs rapports toutes valeurs possibles d'une extrémité à l'autre de l'échelle des rangs : si toutefois on craint des rapports très écartés de 1 dans le cas des angles voisins de $\pi/2$, on pourra restreindre l'analyse e.g. aux 3 000 sommets dont le rang moyen est le plus faible (i.e. éliminer les 387 sommets de rang moyen élevé, pour lesquels les angles approchent $\pi/2$: du point de vue physique ces sommets d'acuité très faible, sont justement les moins remarquables).

Mais en général il semble possible de corriger les fonctions QIV (I, JV) par un coefficient fonction de MR (I), choisi de telle sorte qu'à l'intérieur de chaque classe de la variable MR, les


QT07
QU09
QA09
QL03
QT08
QB02
QL02
QT09
QT10
QU10
QA10
QB01
QL01
QB04
QL04
QU08
QA07
QT06
MR10
QB05
QU07
QA06
QT05
QL05
QU06
MR09
QL06
QU05
QB06
QA05
QT04
MR08
MR07
QL07
MR05
MR04
MR02
MR03
MR06
MR01
QB07
QU04
QA04
QL08
QT03
QU01
QU03
QB08
QA03
QL09
QU02
QT02
QB09
QA02
QL10
QA01

Figure 5

quotients aient même dispersion. on poserait donc :

$$QCIV(I,JV) = QIV(I,JV) * COEF(RM(I),JV) ;$$

Le choix du coefficient se fait d'après l'histogramme de QIV au sein de chaque classe de MR ; ou, ce qui revient au même d'après le tableau de Burt (qui donne pour chaque modalité de MR, la distribution des modalités des quotients Q). Du fait qu'on utilise le codage logique, qui diminue grandement l'influence des valeurs extrêmes, ce recodage devrait aboutir à des résultats stables.

Il importe de noter, en conclusion, que même si les transformations suggérées ici ne se font pas sans tatonnements ; le but poursuivi ici qui est de comparer des données physiques expérimentales à des données engendrées par simulation suivant un modèle théorique (ou empirique), sera de toute façon atteint ; l'analyse manifestant nettement les différences de toute sorte entre deux tableaux de données indicernables autrement.