ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 24 — 1991

Approche exploratoire de l'inégalité en situation multivariée

Alain BACCINI, Antoine de FALGUEROLLES, El Mustapha QANNARI *

RÉSUMÉ. – Dans cet article, nous proposons une approche exploratoire pour aborder le problème de l'étude de l'inégalité dans un contexte multivarié. Deux situations sont étudiées : dans la première, plusieurs variables quantitatives sont observées simultanément et l'on réalise une Décomposition en Valeurs Singulières pour mettre en évidence différents éléments synthétiques de l'inégalité ; dans la seconde, on observe une seule variable quantitative en même temps que deux variables catégorielles et l'on utilise une technique voisine de l'Analyse Factorielle des Correspondances pour réaliser un graphique synthétisant l'explication de l'inégalité par les variables catégorielles. Les deux exemples traités sont relatifs aux revenus des ménages français en 1979.

Empirical Multivariate Analysis of Income Inequality

ABSTRACT. – In this article, we consider the problem of assessing the inequality in a multivariate context, and we present an exploratory approach which is implemented in two practical situations. Firstly, several quantitative variables are simultaneously observed; considering an adapted matrix, we then extract, by means of a Singular Value Decomposition, specific summaries of inequality. Secondly, a quantitative variable and two explanatory polytomous variables are jointly observed; we then produce biplots, by means of an adapted Correspondance Analysis, describing how inequality is related to the levels of the explanatory variables.

* A. BACCINI et A. DE FALGUEROLLES : Laboratoire de Statistique et Probabilités, URA-CNRS D0745, 118, route de Narbonne, 31062 Toulouse Cedex. E. M. QANNARI : ENITIAA, La Géraudière, 44072 Nantes Cedex.

1 Introduction

Dans cet article, deux situations ont retenu notre attention. Dans la première, on considère plusieurs variables statistiques quantitatives, se prêtant chacune à une étude d'inégalité, et conjointement observées sur une même population ; on désire alors procéder à une étude globale de l'inégalité de ces variables et, le cas échéant, mettre en évidence des phénomènes de compensation ou d'aggravation. Dans la seconde, on observe simultanément une variable quantitative se prêtant à une étude d'inégalité et deux variables qualitatives ; on cherche alors à apprécier la manière dont ces variables qualitatives expliquent l'inégalité de la variable quantitative.

Les méthodes que nous proposons ici sont assez simples à mettre en œuvre ; de façon schématique, elles consistent à procéder à des Décompositions en Valeurs Singulières de matrices. Il s'ensuit, de façon classique, des représentations graphiques facilitant l'interprétation des résultats. Notre démarche est analogue à celle des méthodes usuelles d'analyse descriptive des données multidimensionnelles ; elle est donc essentiellement exploratoire.

Ce travail comprend trois parties. Dans le paragraphe 2, nous faisons quelques rappels sur les mesures d'inégalité, nous indiquons deux problèmes fréquemment étudiés dans le contexte considéré, et nous situons les méthodes proposées par rapport à ces problèmes. Les paragraphes 3 et 4 sont consacrés à l'étude des deux situations évoquées ci-dessus, un exemple illustrant chacune d'entre elles.

2 Préliminaires sur l'étude de l'inégalité

2.1. Rappels sur les mesures d'inégalité dans le cas unidimensionnel

On considère une variable statistique quantitative, notée X, observée sur un ensemble fini Ω de n individus (supposés équipondérés dans ce paragraphe). L'individu (ou unité statistique) générique est noté ω. $X(\omega)$ désigne l'observation de la variable X sur l'individu ω.

On suppose que la variable X se prête à une étude d'inégalité (X représente, par exemple, une quantité telle que le revenu, la fortune...) ; cela

revient à supposer que $M_x = \sum_{\omega \in \Omega} X(\omega)$ est une quantité strictement positive et a une signification concrète ; nous appellerons masse totale des valeurs de X cette quantité M_x. Quelquefois, on suppose également que X est à valeurs non négatives, mais cela n'est pas nécessaire ici, comme nous le verrons au paragraphe 3 (voir également la remarque 2).

L'étude de l'inégalité de la distribution de X peut alors être présentée comme suit : on considère deux séries de poids sur les individus, celle de poids *a priori*, $p(\omega) = \frac{1}{n}$, $\forall \omega \in \Omega$, et celle des poids induits par X, $q(\omega) = \frac{X(\omega)}{M_x}$, $q(\omega)$ représentant la part de la masse totale des valeurs de X détenue par l'individu ω ; on procède alors à la comparaison de ces deux séries de poids.

De façon plus précise, la variable X a une distribution d'autant plus inégalitaire que ces deux séries (ou encore les deux fonctions de répartition empiriques respectivement associées F et G) diffèrent ; la plupart des indices synthétiques d'inégalité peuvent être interprétés comme des mesures de la dissemblance soit entre les deux séries de poids, soit entre les deux fonctions de répartition empiriques F et G. Pour plus de détails sur cette approche, on pourra se reporter à Baccini *et al.* [1986].

La technique la plus usitée dans l'étude empirique de l'inégalité est la courbe de Lorenz ; elle consiste, dans un repère plan orthonormé, à joindre les points de coordonnées $(G(i), F(i))$, l'indice i variant de 1 à n et correspondant aux individus de Ω rangés par ordre croissant de valeurs de X (voir le graphique 1 pour des exemples). Il s'agit en fait du « P-P plot » de G et F.

En ce qui concerne les indices synthétiques d'inégalité, ils sont nombreux dans la littérature statistique ; le plus courant est l'indice de Gini, égal à deux fois l'aire de concentration, aire comprise entre la courbe de Lorenz et la première bissectrice (l'indice de Gini est donc compris entre 0 et 1 pour les variables à valeurs non négatives). Nous utiliserons ici le carré du coefficient de variation, rapport de la variance empirique au carré de la moyenne arithmétique :

$$\tau_x^2 = \frac{\sigma_x^2}{\bar{x}^2} = \frac{1}{n\bar{x}^2} \sum_{\omega \in \Omega} [X(\omega) - \bar{x}]^2.$$

Outre la cohérence de cet indice avec l'approche envisagée ici (voir les paragraphes 3 et 4), on sait qu'il est un des seuls (avec l'indice de Theil et l'écart-moyen des logarithmes) à se décomposer de façon satisfaisante sur toute partition de la population étudiée (voir le paragraphe 4.1).

Remarque 1 : La série des poids *a priori* ne correspond pas nécessairement à l'équipondération des individus ; ainsi, certaines situations concrètes peuvent conduire à affecter à chaque individu ω un poids $p(\omega)$ vérifiant :

$$\forall \omega \in \Omega, \qquad p(\omega) > 0 ; \qquad \sum_{\omega \in \Omega} p(\omega) = 1.$$

Le problème du choix de ces poids et des définitions correspondantes de M_x et de $q(\omega)$ est traité en annexe 1.

REMARQUE 2 : Dans le cas très particulier où certaines valeurs $X(\omega)$ sont négatives, les « poids » correspondants $q(\omega)$ sont aussi négatifs ; $\{q(\omega), \omega \in \Omega\}$ n'est plus alors une loi de probabilité, mais une mesure, sur $(\Omega, \mathscr{P}(\Omega))$. Bien qu'un peu inhabituelle, cette situation n'est pas génante, puisque toutes les propriétés des techniques présentées ici sont conservées ; de plus, il se trouve que c'est une situation que l'on rencontre parfois dans la pratique, comme nous le verrons avec l'exemple traité en 3.6.

REMARQUE 3 : Les poids $q(\omega)$ sont invariants dans toute homothétie de X de rapport positif. Autrement dit, les études d'inégalité sur X sont indépendantes de l'unité dans laquelle est exprimée X. Il est donc possible de supposer, sans perte de généralité, que X est de moyenne égale à 1 ; le carré du coefficient de variation est alors confondu avec la variance.

2.2. Deux problèmes classiques dans l'étude de l'inégalité

De nombreux articles consacrés aux mesures de l'inégalité abordent l'un ou l'autre des problèmes suivants :

- La variable X est définie comme somme d'un certain nombre de composantes, chacune pouvant faire l'objet d'une étude d'inégalité (par exemple, le revenu des ménages décomposé selon différentes sources) ; le problème est alors d'exprimer l'inégalité totale en fonction de l'inégalité de chaque composante ; diverses réponses ont été fournies par RAO [1969], KAKWANI [1977], FEI *et al.* [1978], PYATT *et al.* [1980], et SHORROCKS [1982 et 1983] ;
- une partition en K classes est définie sur Ω (par exemple, à partir d'une variable qualitative à K modalités) ; si l'on mesure l'inégalité partielle de la distribution de X sur chaque classe de la partition, le problème est dans ce cas de retrouver l'inégalité totale en fonction de ces inégalités partielles, et d'étudier les indices synthétiques qui se prêtent le mieux à cette décomposition ; ce problème est abordé par RAO [1969], BOURGUIGNON [1979], SHORROCKS [1980], PYATT *et al.* [1980], DAS et PARIKH [1982], SHORROCKS [1984] et BOURGUIGNON et MORRISSON [1985].

L'objet du travail présenté ici n'est pas de répondre directement aux problèmes mentionnés ci-dessus (au demeurant, les différents articles cités semblent avoir bien fait le tour de la question), mais plutôt de montrer comment une approche exploratoire, s'appuyant sur des techniques classiques d'analyse des données multidimensionnelles, peut apporter un complément d'information face à une situation concrète relevant de ces problèmes.

2.3. Principes de l'approche multidimensionnelle

Partant du fait que la plupart des mesures d'inégalité découlent, dans le cas unidimensionnel, de la comparaison de deux séries de poids, nous

proposons ici d'utiliser une technique de comparaison de tableaux pour généraliser cette approche dans un contexte multidimensionnel.

Le problème de la comparaison de tableaux dans le cadre de la statistique multidimensionnelle a fait l'objet, ces dernières années, de diverses études ; nous nous référons plus particulièrement ici à DOMENGES et VOLLE [1979], QANNARI [1983], ESCOFIER [1984], CAUSSINUS et FALGUEROLLES [1987], FALGUEROLLES et HEIJDEN [1987] et HEIJDEN [1987].

Notre approche est en fait basée sur l'Analyse en Composantes Principales (ACP) d'un tableau, ou, ce qui est équivalent, à sa Décomposition en Valeurs Singulières (DVS).

3 Étude de l'inégalité conjointe de plusieurs variables quantitatives

3.1. Problème

On considère ici p ($p \geqslant 2$) variables statistiques quantitatives notées X^j, $j = 1, \ldots, p$, simultanément observées sur une population Ω (de cardinal n) dans laquelle chaque individu ω est affecté d'un poids quelconque $p(\omega)$. On suppose que chacune de ces variables se prête séparément à une étude d'inégalité, et l'on souhaite étudier l'inégalité globale engendrée par les p variables prises conjointement. En particulier, l'idée directrice est de mettre en évidence les variables qui renforcent ou compensent les phénomènes d'inégalité.

REMARQUE 4 : La somme des variables $\sum_{j=1}^{p} X^j$ n'a pas nécessairement de signification concrète. Elle peut en avoir une, comme nous le verrons dans l'exemple présenté en 3.6 (somme des revenus des ménages provenant de diverses sources), mais ce n'est pas nécessaire dans ce qui suit.

Comme vu précédemment, chaque variable X^j induit un système de poids q^j, où $q^j(\omega)$ est la part relative de la masse totale des valeurs de X^j détenue par l'individu ω.

Convention : Nous supposerons dorénavant que la moyenne arithmétique de chaque variable X^j vaut 1 (voir remarque 3). De plus, si les poids $p(\omega)$ ne sont pas tous identiques, nous supposerons que l'on a affaire soit à des données brutes, soit à des données « moyennées » (voir annexe 1).

Les deux hypothèses formulées ci-dessus permettent de déduire d'une part que le carré du coefficient de variation de chaque variable X^j est égal à sa variance, d'autre part que les poids induits par les variables X^j ont pour

expression :

$$q^j(\omega) = p(\omega)\,X^j(\omega), \qquad \forall j = 1, \ldots, p \quad \text{et} \quad \forall \omega \in \Omega.$$

3.2. Méthode

La méthode proposée s'inspire de l'approche suggérée par DOMENGE et VOLLE [1979], et utilisée par QANNARI [1983] pour analyser une famille de mesures (ici les $q^j(\omega)$, $j = 1, \ldots, p$) en référence à une probabilité [ici $p(\omega)$]. On construit donc le tableau T à n lignes et p colonnes d'élément général :

$$T^j(\omega) = \frac{q^j(\omega) - p(\omega)}{\sqrt{p(\omega)}}, \qquad j = 1, \ldots, p \quad \text{et} \quad \omega = 1, \ldots, n.$$

Puisque $q^j(\omega) = p(\omega)\,X^j(\omega)$, il vient encore $T^j(\omega) = \sqrt{p(\omega)}\,[X^j(\omega) - 1]$, où $[X^j(\omega) - 1]$ mesure l'écart entre la valeur de X^j détenue par l'individu ω et la moyenne de cette variable; c'est même un écart relatif puisque les variables sont préalablement divisées par leur moyenne.

De la définition de T, on déduit la propriété suivante :

> PROPRIÉTÉ 1 : La trace de ${}^tTT$ est égale à la somme des carrés des coefficients de variation des variables X^j, $j = 1, \ldots, p$ (c'est aussi la somme des carrés des coefficients de variation des variables avant transformation).

Il suffit en effet de remarquer que

$$\sum_{\omega \in \Omega} T^j(\omega)\,T^k(\omega) = \sum_{\omega \in \Omega} p(\omega)\,[X^j(\omega) - 1]\,[X^k(\omega) - 1].$$

${}^tTT$ est donc la matrice des variances-covariances des variables X^j, $j = 1, \ldots, p$.

La propriété ci-dessus suggère de procéder à la Décomposition en Valeurs Singulières de T, pour obtenir les composantes principales associées aux valeurs propres de ${}^tTT$ (voir annexe 2). Les propriétés de ces composantes principales, ainsi que leur interprétation, permettent de définir des variables d'inégalité principale.

3.3. Variables d'inégalité principale

On réalise donc la DVS du tableau T, ce qui est équivalent à l'ACP de T non centrée, non réduite, les poids des lignes valant chacun 1 (le poids des individus est pris en compte dans T), et la métrique dans l'espace des individus ($\mathbb{R}^p$) étant l'identité.

Désignons par $C^1, \ldots, C^p$ les composantes principales obtenues. Par analogie avec les données initiales dont elles sont issues, il est naturel de

considérer que ces composantes sont de la forme

$$C^k(\omega) = \frac{\mu^k(\omega) - p(\omega)}{\sqrt{p(\omega)}}, \qquad k = 1, \ldots, p \quad \text{et} \quad \omega = 1, \ldots, n,$$

$\mu^k(\omega)$ désignant la part $p(\omega)\Gamma^k(\omega)$ de la masse totale des valeurs d'une variable Γ^k (supposée sans dimension et de moyenne 1) détenue par l'individu ω.

Nous sommes donc amenés à donner la définition suivante :

Définition : La variable Γ^k, définie par $\Gamma^k(\omega) = \dfrac{C^k(\omega)}{\sqrt{p(\omega)}} + 1, \quad \forall \omega \in \Omega,$ est appelée k-ième variable d'inégalité principale.

Ainsi définies, les variables Γ^k possèdent un certain nombre de propriétés, « naturelles » dans le contexte de l'ACP. Pour les établir, nous avons besoin du lemme qui suit.

LEMME 2 : $\sum_{\omega \in \Omega} \sqrt{p(\omega)}\, C^k(\omega) = 0, \ \forall k = 1, \ldots, p.$

Il est en effet immédiat de vérifier que

$$\sum_{\omega \in \Omega} \sqrt{p(\omega)}\, T^j(\omega) = \sum_{\omega \in \Omega} p(\omega)[X^j(\omega) - 1] = 0, \ \forall j = 1, \ldots, p;$$

par suite, chaque C^k, combinaison linéaire des T^j, vérifie également cette propriété.

PROPRIÉTÉ 3 : Γ^k, k-ième variable d'inégalité principale, est de moyenne 1.

Ceci découle immédiatement du lemme précédent. On remarquera qu'une autre façon d'énoncer ce résultat consiste à écrire $\sum_{\omega \in \Omega} \mu^k(\omega) = 1$, ce qui rend cohérent la définition des parts $\mu^k(\omega)$.

PROPRIÉTÉ 4 : Le coefficient de variation de Γ^k, τ_k, est égal à la k-ième valeur singulière s_k de T (dans le rangement par ordre décroissant de ces valeurs). Autrement dit, $\tau_k^2 = \lambda_k$, k-ième valeur propre de l'ACP de T.

En effet :

$$\tau_k^2 = \mathrm{Var}(\Gamma^k) = \sum_{\omega \in \Omega} p(\omega)[\Gamma^k(\omega) - 1]^2$$
$$= \sum_{\omega \in \Omega} [C^k(\omega)]^2 = \lambda_k = s_k^2 \text{ (voir annexe 2).}$$

PROPRIÉTÉ 5 : Les variables d'inégalité principale sont deux à deux non corrélées.

Ce résultat se vérifie de la même façon que le précédent.

PROPRIÉTÉ 6 : $X^j (j = 1, \ldots, p)$ et $\Gamma^k (k = 1, \ldots, p)$ sont de covariance égale au produit scalaire usuel entre la j-ième colonne de T et la k-ième composante principale.

En effet :

$$\begin{aligned}\mathrm{Cov}(X^j, \Gamma^k) &= \sum_{\omega \in \Omega} p(\omega)[X^j(\omega) - 1][\Gamma^k(\omega) - 1] \\ &= \sum_{\omega \in \Omega} C^k(\omega) \frac{p(\omega) X^j(\omega) - p(\omega)}{\sqrt{p(\omega)}} \\ &= \sum_{\omega \in \Omega} C^k(\omega) T^j(\omega).\end{aligned}$$

Ce résultat est utilisé dans l'interprétation des graphiques relatifs aux variables.

REMARQUE 5 : Les variables d'inégalité principale peuvent présenter certaines valeurs négatives, même lorsque toutes les variables initiales sont à valeurs positives.

REMARQUE 6 : Il est bien connu que si C^k est la k-ième composante principale, $-C^k$ peut aussi être retenue comme composante principale de même rang. Ce choix étant arbitraire, il importe d'étudier comment cette indétermination affecte la construction de la k-ième variable d'inégalité principale. On vérifie aisément que les déterminations Γ^k et Γ_0^k de cette variable, respectivement associées à C^k et $-C^k$, sont liées par la relation $\Gamma^k(\omega) + \Gamma_0^k(\omega) = 2$, $\forall \omega \in \Omega$. Il en résulte que Γ^k et Γ_0^k, de même moyenne et de même variance, présentent les mêmes valeurs d'indices synthétiques usuels d'inégalité (coefficient de variation, indice de GINI...). Dans la pratique, pour choisir entre Γ^k et Γ_0^k, on pourra utiliser leurs corrélations avec les variables initiales. On notera encore que la courbe de LORENZ de Γ_0^k n'est pas confondue avec la « pseudo courbe de LORENZ » de Γ^k obtenue pour l'ordre induit par ses valeurs décroissantes (voir FEI *et al.* [1978]). Toutefois, les relations entre ces diverses courbes sont simples à expliciter.

REMARQUE 7 : Il est immédiat de vérifier que la méthode proposée est invariante si l'on modifie l'ordre d'observation des individus ou encore si l'on multiplie toute variable X^j par une constante positive quelconque. Par contre, ce n'est plus le cas si l'on rajoute une constante non nulle à une variable X^j.

3.4. Représentations graphiques

Selon le point de vue couramment adopté en analyse exploratoire des données, on dispose de deux types de représentations graphiques ; le premier « visualise » les individus, le second les variables.

- *Graphiques relatifs aux individus*

Deux variables d'inégalité principale Γ^k et Γ^l étant retenues [le plus souvent, $(k, l) = (1, 2)$ ou $(1, 3)$], chaque individu ω, $\omega = 1, \ldots, n$, est représenté dans un repère plan orthonormé par le point de coordonnées $(\Gamma^k(\omega) - 1, \Gamma^l(\omega) - 1)$. En effet, $\Gamma^k(\omega) - 1$ (resp. $\Gamma^l(\omega) - 1$) s'interprète comme la différence relative entre la valeur de Γ^k (resp. de Γ^l) présentée

par l'individu ω et la moyenne de Γ^k (resp. de Γ^l), puisque cette moyenne est égale à 1.

Dans la pratique, ces graphiques ne seront intéressants que dans le cas où n sera suffisamment petit ; toutefois, dans les autres cas, on pourra encore les utiliser dans une perspective illustrative, par exemple pour représenter les barycentres des individus appartenant à des sous-populations définies sur Ω.

REMARQUE 8 : Il est facile de voir que lorsque $p(\omega)=\frac{1}{n}$, $\forall \omega \in \Omega$, ces représentations graphiques sont directement fournies par un logiciel d'Analyse en Composantes Principales. En effet, si l'on effectue une ACP du tableau $T\left(T^j(\omega)=\frac{1}{\sqrt{n}}[X^j(\omega)-1]\right)$, les composantes principales C^k sont centrées (puisque les colonnes de T le sont dans ce cas), et l'on a $\Gamma^k(\omega)-1=\sqrt{n}\,C^k(\omega)$. Seule l'échelle du graphique fourni est à modifier pour tenir compte du coefficient $\sqrt{n}$. Par contre, lorsque les $p(\omega)$ ne sont pas tous égaux, une ACP non centrée du tableau T fournit les composantes principales $C^k(\omega)$ et l'on doit alors calculer les valeurs $\Gamma^k(\omega)-1=\frac{C^k(\omega)}{\sqrt{p(\omega)}}$.

• *Graphiques relatifs aux variables*

La DVS de T fournit une base orthonormée du sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de T. La coordonnée de la j-ième colonne T^j de T sur le k-ième élément de cette base (e^k) est donnée par la relation classique suivante :

$$\frac{\langle T^j, C^k \rangle}{\|C^k\|} = \frac{\sum_{\omega \in \Omega} T^j(\omega)\, C^k(\omega)}{\left\{\sum_{\omega \in \Omega} [C^k(\omega)]^2\right\}^{1/2}}.$$

De même, le cosinus de l'angle θ_j^k entre T^j et e^k s'écrit :

$$\cos\theta_j^k = \frac{\sum_{\omega \in \Omega} T^j(\omega)\, C^k(\omega)}{\left\{\sum_{\omega \in \Omega} [T^j(\omega)]^2\right\}^{1/2} \left\{\sum_{\omega \in \Omega} [C^k(\omega)]^2\right\}^{1/2}}.$$

On peut vérifier sans difficulté que $\sum_{\omega \in \Omega} [T^j(\omega)]^2 = \mathrm{Var}(X^j)$; compte-tenu des propriétés 4 et 6, on voit donc que $\cos\theta_j^k$ représente le coefficient de corrélation linéaire entre la variable X^j et la k-ième variable d'inégalité principale Γ^k.

Par suite, si l'on effectue une ACP non centrée du tableau T, les sorties graphiques concernant les « variables » sont interprétables de la même façon que dans une ACP usuelle.

On peut aussi, comme dans une ACP usuelle, représenter des variables supplémentaires (ou illustratives) qui ne sont pas intervenues dans la réalisation de l'analyse, mais qui présentent un intérêt particulier (par exemple, le revenu total dans l'exemple traité en 3.6).

3.5. Utilisation de l'ensemble des variables d'inégalité principale

Contrairement à la pratique courante de l'ACP, on s'intéresse ici à la décomposition spectrale complète de la matrice ${}^tTT$. En effet, si les premières composantes décrivent les « inégalités maximales », les dernières décrivent les « inégalités minimales » et peuvent, dans une certaine optique, être tout aussi intéressantes. D'où l'intérêt de considérer le tableau complet des corrélations entre variables initiales et variables d'inégalité principale.

3.6. Exemple

Les données présentées ici, ainsi que celles du paragraphe 4.5, sont extraites de l'enquête INSEE sur les revenus fiscaux des ménages français en 1979[1] ; on pourra trouver une présentation générale de ces données dans CANCEILL [1984].

Nous avons considéré ici un échantillon aléatoire de 1371 ménages fiscaux de la région MIDI-PYRENEES ($n=1371$) ; bien qu'il s'agisse d'un cas dans lequel le redressement d'échantillon se trouve pleinement justifié (voir CANCEILL, [1984]), nous avons, pour simplifier les choses, utilisé l'équipondération des ménages, notre démarche étant ici essentiellement illustrative. 6 variables quantitatives ont été prises en compte ($p=6$), chacune représentant une source particulière de revenus :

X^1 : revenus du chef de ménage (pour certaines professions non salariées, les valeurs de cette variable peuvent être négatives ou nulles) ;

X^2 : revenus du conjoint du chef de ménage (quelques observations de cette variable sont négatives, beaucoup sont nulles) ;

X^3 : revenus des autres membres du ménage (la plupart des valeurs de X^3 sont nulles) ;

X^4 : revenus non individualisables (même remarque que pour X^2) ;

X^5 : prestations familiales non soumises à conditions de ressources (même remarque que pour X^3) ;

X^6 : prestations familiales soumises à conditions de ressources (même remarque).

---

1. Nous tenons à remercier Madame G. CANCEILL, alors au département « Population-Ménages » de l'INSEE, de nous avoir fourni ces données.

Nous avons, de plus, calculé deux autres variables :

$X^7 = X^5 + X^6$ (total des prestations familiales) ;

$X^8 = \sum_{j=1}^{6} X^j$ (total des revenus du ménage).

Le tableau 1 présente les statistiques élémentaires sur ces données.

TABLEAU 1

• *Résumés numériques des variables (en francs, arrondis à la centaine près)*

	Minimum	Maximum	Moyenne	Ecart-type
X^1	− 32 300	1 885 700	58 300	84 600
X^2	− 4 800	289 400	10 800	22 300
X^3	0	231 300	6 000	16 600
X^4	− 113 100	1 041 800	5 700	44 900
X^5	0	32 800	1 600	3 800
X^6	0	16 200	1 000	2 400

• *Matrice des corrélations*

	X^1	X^2	X^3	X^4	X^5	X^6
X^1	1	0,08	0,03	0,14	0,06	− 0,13
X^2	*	1	− 0,05	0,02	0,03	− 0,10
X^3	*	*	1	− 0,01	− 0,04	− 0,04
X^4	*	*	*	1	− 0,03	− 0,04
X^5	*	*	*	*	1	0,64
X^6	*	*	*	*	*	1

A titre d'illustration, nous avons, dans le graphique 1, tracé la courbe de LORENZ de chacune des 8 variables et indiqué les valeurs du coefficient de GINI et du carré du coefficient de variation.

Nous avons alors réalisé la DVS du tableau T (1371 × 6) correspondant à ces données. Nous en donnons dans le tableau 2 les principaux résultats.

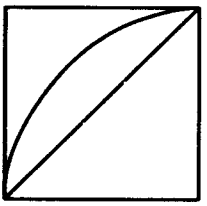
On notera que, dans l'exemple, les représentations graphiques présentées en 3.4 ne sont pas d'une grande utilité. En effet, le graphique des individus n'est que de peu d'intérêt, puisqu'il y en a 1371, et le graphique des variables n'apporte rien par rapport au tableau des corrélations entre les variables initiales et les variables d'inégalité principale, dans la mesure où ce dernier est ici très clair. On constate en effet que chaque variable d'inégalité principale représente soit une seule variable initiale, soit deux (dans le cas de Γ^2, représentant X^5 et X^6) ; ceci est bien entendu lié au fait que les variables initiales sont pratiquement non corrélées, à l'exception de X^5 et X^6.

On voit donc, à travers cet exemple, que l'un des principaux intérêts de la méthode est de mettre en évidence la structuration des données initiales selon 5 « directions » d'inégalité, chacune représentant une variable initiale ou un groupe de ces variables ; l'inégalité présentée par chaque « direction » (ou variable d'inégalité principale) est mesurée par le carré du coefficient

GRAPHIQUE 1

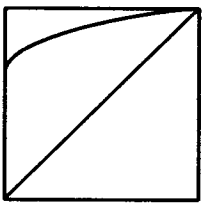
Courbes de LORENZ des variables X^1 à X^8 et valeurs du coefficient de GINI et du carré du coefficient de variation

X^1



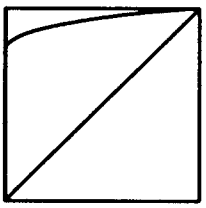
$G(X^1) = 0,51$
$\tau^2(X^1) = 2,11$

X^2



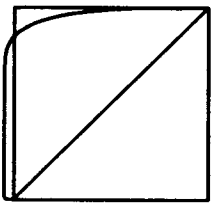
$G(X^2) = 0,81$
$\tau^2(X^2) = 4,27$

X^3



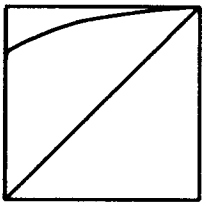
$G(X^3) = 0,88$
$\tau^2(X^3) = 7,73$

X^4



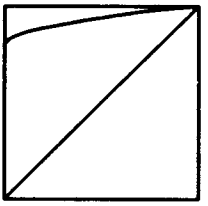
$G(X^4) = 1,02$
$\tau^2(X^4) = 61,16$

X^5



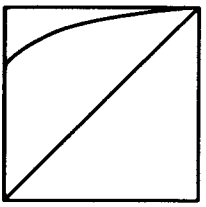
$G(X^5) = 0,85$
$\tau^2(X^5) = 5,50$

X^6



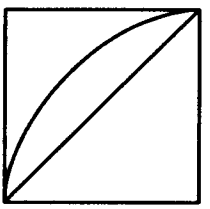
$G(X^6) = 0,87$
$\tau^2(X^6) = 5,79$

X^7



$G(X^7) = 0,82$
$\tau^2(X^7) = 4,67$

X^8



$G(X^8) = 0,47$
$\tau^2(X^8) = 1,64$

TABLEAU 2

(*a*) Carrés des valeurs singulières de la DVS (ou valeurs propres de l'ACP) :

$\lambda_1 = 61{,}21 \quad \lambda_2 = 9{,}34 \quad \lambda_3 = 7{,}66 \quad \lambda_4 = 4{,}36 \quad \lambda_5 = 2{,}40 \quad \lambda_6 = 1{,}57$

(*b*) Pourcentages correspondants :

$p_1 = 70{,}73 \quad p_2 = 10{,}79 \quad p_3 = 8{,}86 \quad p_4 = 5{,}04 \quad p_5 = 2{,}77 \quad p_6 = 1{,}82$

(*c*) Pourcentages cumulés :

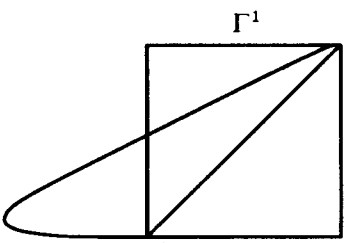
$pc_1 = 70{,}73 \quad pc_2 = 81{,}52 \quad pc_3 = 90{,}38 \quad pc_4 = 95{,}41 \quad pc_5 = 98{,}18 \quad pc_6 = 100.$

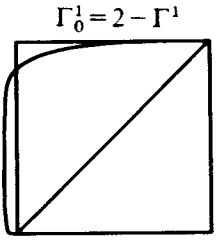
(*d*) Corrélations entre les variables initiales (X^j) et les variables d'inégalité principale (Γ^k) :

	Γ^1	Γ^2	Γ^3	Γ^4	Γ^5	Γ^6
X^1	−0,14	−0,06	0,03	0,21	0,77	0,58
X^2	−0,02	−0,05	−0,13	0,97	−0,19	0,04
X^3	0,01	−0,26	0,96	0,06	−0,02	−0,01
X^4	−1,00	0,01	0,00	0,00	0,00	0,00
X^5	0,04	0,87	0,18	0,17	0,31	−0,28
X^6	0,04	0,89	0,20	−0,09	−0,29	0,27

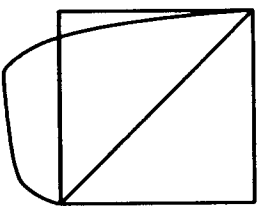
GRAPHIQUE 2

Courbes de LORENZ et principales caractéristiques numériques des deux premières variables d'inégalité principale





$\bar{\gamma}_1 = 1 \; ; \; \tau(\Gamma^1) = \sigma(\Gamma^1) = 7{,}82 \; ;$
$G(\Gamma^1) = 1{,}05 \; ; \; \tau^2(\Gamma^1) = \sigma^2(\Gamma^1) = 61{,}21.$



$\bar{\gamma}_2 = 1 \; ; \; \tau(\Gamma^2) = \sigma(\Gamma^2) = 3{,}06 \; ;$
$G(\Gamma^2) = 1{,}25 \; ; \; \tau^2(\Gamma^2) = \sigma^2(\Gamma^2) = 9{,}34.$

de variation correspondant et, donc, tient compte des autres variables. Ainsi, on constate que l'ensemble des prestations familiales demeure moins inégalitaire que les revenus non individualisables, mais davantage que les revenus des autres personnes du ménage.

Précisons que le choix entre Γ^k et Γ_0^k se fait ici de façon très simple : on choisit celle de ces deux variables ayant une corrélation positive avec la (les) variable(s) initiale(s) qu'elle représente. Nous donnons dans le graphique 2 les courbes de LORENZ de Γ^1, de Γ_0^1 et de Γ^2.

On peut remarquer que Γ^2, tout en ayant un coefficient de variation nettement plus petit que Γ^1 ou Γ_0^1 ($\tau_1^2 = 61{,}21$ et $\tau_2^2 = 9{,}34$), a un coefficient de GINI supérieur ($G(\Gamma^1) = 1{,}05$ et $G(\Gamma^2) = 1{,}25$). Cela est possible dans la mesure où le coefficient GINI n'est pas directement lié à la variance, et s'explique ici par le fait que la grande variance de Γ^1 est surtout due à des valeurs extrêmes (en particulier, l'intervalle interquartiles est de 0,16 pour Γ^1 et de 1,41 pour Γ^2).

Signalons, pour terminer cet exemple, que l'Analyse en Composantes Principales usuelle ne donne pas des résultats très intéressants sur ces données : dans l'ACP non réduite, chacun des quatre premiers facteurs représente une des quatre premières variables initiales (classées par ordre décroissant de variance) et le cinquième représente X^5 et X^6 ; dans l'ACP réduite, le premier facteur représente X^5 et X^6 et l'interprétation des autres est très confuse.

4 Analyse de l'inégalité d'une variable quantitative par rapport à deux variables qualitatives

4.1. Introduction

Nous avons déjà signalé en 2.2 que l'étude de l'inégalité d'une variable quantitative, au sein d'une population dans laquelle est définie une partition, a fait l'objet, depuis quelques années, d'un certain nombre d'articles. L'objectif essentiel de ces articles est de déterminer quelles mesures d'inégalité ont, dans les conditions indiquées ci-dessus, des « propriétés intéressantes ».

Ces propriétés sont, d'une part, des propriétés générales, qu'il est habituel d'imposer à une mesure d'inégalité I :

- symétrie (I est invariante lorsqu'on permute les observations de la variable quantitative),

• homogénéité de degré zéro, c'est-à-dire invariance d'échelle (invariance de I si on multiplie par une constante positive toutes les valeurs de la variable),

• éventuellement, diverses conditions de régularité (continuité de I, existence de dérivées partielles...).

D'autre part s'ajoute à cela la propriété fondamentale de décomposabilité :

$$I = I_B + \sum_{k=1}^{K} \alpha_k I_k,$$

où I_B est l'inégalité entre les classes, I_k l'inégalité à l'intérieur de la k-ième classe et α_k une pondération de cette classe, l'ensemble des pondérations vérifiant $\sum_{k=1}^{K} \alpha_k = 1$.

De toutes les mesures classiques d'inégalité, trois seulement possèdent les propriétés requises : l'indice de THEIL (voir THEIL [1967] : $I_T = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\bar{y}} \text{Log} \frac{y_i}{\bar{y}}$, où $\{y_i;\ i=1, \ldots, n\}$ est l'ensemble des observations de la variable considérée et $\bar{y}$ sa moyenne arithmétique), l'écart-moyen des logarithmes $\left(I_L = \text{Log}\, \bar{y} - \frac{1}{n} \sum_{i=1}^{n} \text{Log}\, y_i \right)$ et le carré du coefficient de variation. Pour tous ces résultats, on pourra se reporter à BOURGUIGNON [1979] et à SHORROCKS [1980 et 1984].

Dans cette partie, nous abordons le problème de la mesure de l'inégalité d'une variable quantitative relativement à deux variables qualitatives. Des trois indices cités ci-dessus, nous avons retenu le dernier, égal à la variance lorsqu'on considère une variable de moyenne 1, et qui, par conséquent, se prête bien à une approche de type Décomposition en Valeurs Singulières ; en effet, c'est encore cette approche qui sera utilisée ici.

Notons que le même problème a déjà été abordé, mais en utilisant l'indice de THEIL et une optique voisine de celle de l'analyse de variance, par BOURGUIGNON et MORRISSON [1985]. Toutefois, contrairement à ce qui est fait dans l'article précité, nous proposons ici une méthode permettant de traiter le cas où l'on ne dispose pas des données intra-groupes et où l'on doit donc se contenter de l'analyse de l'inégalité inter-groupes.

4.2. Notations et objectif

Notons toujours Ω la population considérée, dans laquelle chaque individu est supposé affecté du même poids $\frac{1}{n}$, si $n = \text{card}(\Omega)$. Si nécessaire, la méthode se généralise sans difficulté à des pondérations quelconques (voir annexe 1, cas de données « moyennées »).

On observe donc 3 variables sur Ω; l'une, notée Z, est quantitative, se prête à une étude d'inégalité, et est toujours supposée, sans perte de généralité, de moyenne 1 (la méthode qui suit est en effet invariante dans toute homothétie de rapport positif sur Z); les deux autres sont qualitatives : X (resp. Y) présentant I (resp. J) modalités notées $X_1, \ldots, X_I$ (resp. $Y_1, \ldots, Y_J$).

On cherche alors à apprécier la manière dont les variables X et Y « expliquent », notamment à travers leur interaction, l'inégalité de la variable Z; nous proposons pour cela une approche basée sur les méthodes factorielles.

Notons Ω_{ij} la sous-population de Ω constituée par les unités statistiques présentant conjointement les modalités X_i de X et Y_j de Y, p_{ij} le poids de cette sous-population $\left(p_{ij}=\frac{n_{ij}}{n}, \text{ avec } n_{ij}=\text{card}\,(\Omega_{ij})\right)$, $\bar{z}_{ij}$ la moyenne de la restriction de Z à cette sous-population $\left(\bar{z}_{ij}=\frac{1}{n_{ij}} \sum_{\omega \in \Omega_{ij}} Z(\omega)\right)$ et $q_{ij}=p_{ij}\bar{z}_{ij}$ la part de la masse totale des valeurs de Z détenue par les unités statistiques de Ω_{ij}.

La variance totale de Z, notée σ^2, se décompose en la somme d'une variance expliquée par la partition de Ω en classes Ω_{ij} (variance inter-classes des moyennes $\bar{z}_{ij}$, notée σ_B^2) et d'une variance résiduelle (moyenne des variances intra-classes, notée σ_W^2). Puisque Z est de moyenne 1, σ^2 s'identifie au carré du coefficient de variation τ^2 pour lequel on a donc la formule :

$$\tau^2=\tau_B^2+\tau_W^2.$$

4.3. Méthode

Pour chaque classe Ω_{ij} de Ω, on compare, ici encore, la part q_{ij} de la masse totale des valeurs de Z à la fréquence relative p_{ij}; on construit donc le tableau T, $I \times J$, d'élément général

$$T_i^j=\frac{q_{ij}-p_{ij}}{\sqrt{p_{ij}}}, \qquad i=1, \ldots, I \quad \text{et} \quad j=1, \ldots, J.$$

Il est clair qu'une valeur positive (resp. négative) de T_i^j correspond à des modalités X_i de X et Y_j de Y associées à un phénomène de sur-inégalité (resp. de sous-inégalité) de Z. Autrement dit, la valeur de T_i^j indique si la sous-population Ω_{ij} détient une part de la masse totale de Z supérieure ou inférieure à sa fréquence relative dans Ω.

On peut remarquer que $T_i^j=\sqrt{p_{ij}}(\bar{z}_{ij}-1)$, et l'on déduit immédiatement la propriété qui suit :

PROPRIÉTÉ 7 : Trace $({}^tTT)=\tau_B^2$.

La structure des éléments de T suggère d'adopter une démarche très voisine de celle retenue en Analyse Factorielle des Correspondances (AFC). En effet, étant donnée une table de contingence relative à deux variables qualitatives, l'AFC fournit une représentation simultanée des modalités de

ces deux variables. Cette représentation peut être obtenue en considérant la DVS du tableau Θ d'élément général $\Theta_i^j = \frac{p_{ij} - p_{i+}\, p_{+j}}{\sqrt{p_{i+}\, p_{+j}}}$, avec $p_{i+} = \sum_{j=1}^{J} p_{ij}$ et $p_{+j} = \sum_{i=1}^{I} p_{ij}$, Θ_i^j comparant la fréquence observée p_{ij} à la fréquence théorique, sous l'hypothèse d'indépendance des deux variables, $p_{i+}\, p_{+j}$, et la trace de ${}^t\Theta\Theta$ étant égale au Φ^2 de contingence (pour cette approche de l'AFC, voir la discussion de Gower dans Deville et Malinvaud [1983] ou encore Falguerolles et Heijden [1987]).

Considérons donc la DVS de T, supposée de rang r (voir annexe 2). T admet r valeurs singulières $s_1 \geqslant \ldots \geqslant s_r$, telles que $s_1^2 + \ldots + s_r^2 = \tau_B^2$. En désignant par c^k (resp. d^k), $k = 1, \ldots, r$, les composantes principales des lignes (resp. des colonnes) du tableau T, il vient :

$$T = \sum_{k=1}^{r} \frac{1}{s_k} c^k\, {}^t d^k.$$

Autrement dit, pour chaque croisement d'une modalité de X et d'une modalité de Y, on a :

$$\frac{(q_{ij} - p_{ij})}{\sqrt{p_{ij}}} = \sum_{k=1}^{r} \frac{1}{s_k} c_i^k\, d_j^k.$$

Il est alors possible, à partir de ce résultat, de réaliser un graphique représentant l'état de concentration de Z qui accompagne les différents croisements de modalités de X et de Y. En particulier, on remarque que si les coordonnées $(c_i^1, \ldots, c_i^r)$ de X_i et les coordonnées $(d_j^1, \ldots, d_j^r)$ de Y_j sont deux à deux de même signe (resp. de signes opposés), ces modalités correspondent à une sur-inégalité de Z (resp. une sous-inégalité de Z).

4.4. Représentations graphiques

On approche T en ne retenant qu'un nombre réduit h, $h \leqslant r$, de composantes ; la qualité globale de cette approximation est mesurée par le pourcentage du carré du coefficient de variation inter-classes dont les h premières valeurs singulières rendent compte :

$$100\, \frac{s_1^2 + \ldots + s_h^2}{\tau_B^2}.$$

Comme en AFC, on peut réaliser un graphique pour chaque couple (k, l) de composantes distinctes $(1 \leqslant k < l \leqslant h)$. Dans un repère plan orthonormé, on associe à chaque modalité X_i de X (resp. Y_j de Y) un point de coordonnées (c_i^k, c_i^l) (resp. (d_j^k, d_j^l)).

Les graphiques réalisés permettent alors :

- de dégager des ressemblances ou des dissemblances entre modalités de X (resp. de Y),

• de reconnaître des associations de modalités de X et de Y qui accompagnent des phénomènes de sur-inégalité ou de sous-inégalité.

REMARQUE 9 : Si le choix des quantités c_i^k et d_j^k pour réaliser les représentations graphiques est conforme à la pratique usuelle de l'AFC, il faut noter que d'autres choix sont possibles. En particulier, pour prendre en compte la totalité des termes $\frac{1}{s_k} c^k {}^t d^k$ dans la décomposition de T, on peut utiliser les quantités $\frac{c_i^k}{s_k^\alpha}$ et $\frac{d_j^k}{s_k^{1-\alpha}}$, $\alpha \in [0, 1]$; les choix $\alpha = 0$, $\alpha = \frac{1}{2}$ et $\alpha = 1$ présentent chacun un certain intérêt. Pour des compléments sur ce point, on pourra se reporter à GOWER et DIGBY [1981], à GABRIEL [1981] ou à LEEUW et HEIJDEN [1988].

4.5. Exemple

On a considéré ici l'ensemble Ω de tous les ménages fiscaux français en 1979; ces ménages sont au nombre de $n = 12\,980\,766$. Ils ont été ventilés selon deux variables catégorielles, la catégorie socio-professionnelle (CSP) et l'âge.

La CSP (X) présente 8 modalités ($I = 8$) :

0 : agriculteurs exploitants;

1 : artisans, patrons pêcheurs, petits et gros commerçants, industriels;

2 : professions non commerciales;

3 : cadres supérieurs;

4 : cadres moyens;

5 : employés;

6 : ouvriers qualifiés;

7 : salariés agricoles et ouvriers non qualifiés.

La variable âge (Y) a été découpée en 9 classes ($J = 9$) : moins de 25 ans, de 25 à 29 ans, de 30 à 34 ans, de 35 à 39 ans, de 40 à 44 ans, de 45 à 49 ans, de 50 à 54 ans, de 55 à 59 ans, plus de 60 ans.

La variable quantitative étudiée (Z) est le revenu total des ménages après impôt. Nous ne disposions en fait que des données agrégées (voir annexe 1) au sein des différentes sous-populations définies par le croisement des modalités de la CSP et des classes d'âge (72); dans chacune de ces sous-populations Ω_{ij}, les données sont l'effectif (n_{ij}) et le cumul des revenus ($\zeta_{ij} = \sum_{\omega \in \Omega_{ij}} Z(\omega) = n_{ij} \bar{z}_{ij}$).

Le tableau 3 présente, pour chaque Ω_{ij}, le revenu moyen $\bar{z}_{ij} = \frac{\zeta_{ij}}{n_{ij}}$; il donne également le revenu moyen par CSP et par classe d'âge, le revenu moyen global, la variance inter-classes du revenu et le carré du coefficient de variation inter-classes.

TABLEAU 3

Revenu moyen des ménages selon la CSP et l'âge (en francs, arrondis à la centaine près)

CSP	Age									
	<25	25-29	30-34	35-39	40-44	45-49	50-54	55-59	>60	
0	39 200	37 600	42 100	44 200	42 900	39 700	38 800	34 900	42 000	39 700
1.	49 100	51 300	73 500	81 500	93 200	94 600	96 900	84 000	98 000	86 800
2.	124 800	50 300	124 600	146 100	143 500	143 300	174 700	127 900	102 200	124 700
3.	30 500	77 500	99 900	113 200	127 400	140 000	132 600	139 600	216 500	123 300
4.	42 700	62 700	76 100	84 000	87 000	88 900	89 200	84 900	81 900	78 600
5.	37 500	54 000	60 000	60 400	64 100	64 100	64 800	59 500	50 600	58 100
6.	44 300	54 100	56 600	55 500	56 700	59 700	61 200	55 200	52 400	55 900
7.	39 500	45 300	50 400	52 400	50 400	53 600	54 200	46 800	40 500	49 500
	41 000	55 200	67 000	72 400	72 500	73 300	73 200	68 700	75 300	67 800

Revenu moyen de l'ensemble des ménages : $\bar{z} = 67\,800$ F ; variance inter-classes de la variable revenu : $\sigma_B^2 = 0{,}665 \times 10^9$; carré de coefficient de variation inter-classes de la variable revenu : $\tau_B^2 = \dfrac{\sigma_B^2}{\bar{z}^2} = 0{,}145$.

Nous avons alors déterminé la matrice T (8×9) de terme général

$$T_i^j = \sqrt{\frac{n_{ij}}{n}}\left(\frac{\bar{z}_{ij}}{\bar{z}} - 1\right) = \sqrt{\frac{n_{ij}}{n}}(\bar{z}'_{ij} - 1),$$

où $\bar{z} = \dfrac{\sum_{i=1}^{I}\sum_{j=1}^{J} \zeta_{ij}}{n}$ et où Z' désigne la variable normalisée (de moyenne 1) déduite de Z. Nous avons ensuite réalisé la DVS de T dont le tableau 4 donne les principaux résultats.

TABLEAU 4

Rang (k)	Valeurs singulières de T (s_k)	Valeurs propres de 'TT (s_k^2)	Pourcentages $\left(100\dfrac{s_k^2}{\tau_B^2}\right)$	Pourcentages cumulés
1	0,354	0,125	86,5	86,5
2	0,118	0,014	9,7	96,2
3	0,060	0,004	2,5	98,7
4	0,033	0,001	0,8	99,5
5	0,022	$0{,}5 \times 10^{-3}$	0,3	99,8
6	0,018	$0{,}3 \times 10^{-3}$	0,2	100
7	0,006	$0{,}3 \times 10^{-4}$	ε_1	100
8	0,002	$0{,}4 \times 10^{-5}$	ε_2	100
		$\sum_{k=1}^{8} s_k^2 = \tau_B^2 = 0{,}145$		

On constate que les deux premières dimensions restituent à elles seules plus de 96 % du carré du coefficient de variation inter-classes. La représentation graphique des modalités des deux variables, réalisée selon le principe indiqué en 4.4 par rapport aux axes 1 et 2, reflète donc presque parfaitement la réalité des inégalités des revenus selon les CSP et les âges. Cette représentation est donnée par le graphique 3.

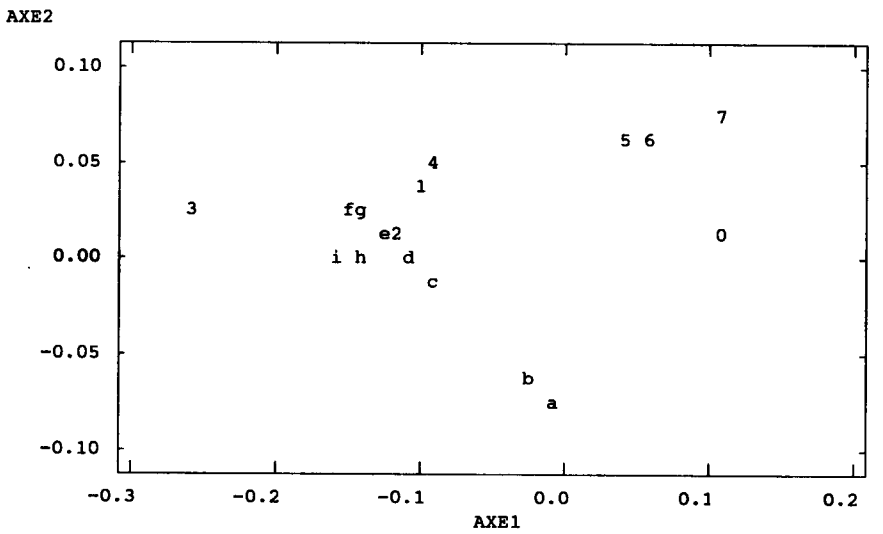
En ce qui concerne les CSP, on remarque essentiellement l'opposition très nette entre les CSP aux revenus les plus faibles (0, 5, 6 et 7), présentant des valeurs positives sur l'axe 1, et celles aux revenus les plus élevés (1, 2, 3 et 4), présentant des valeurs négatives.

En ce qui concerne l'âge, on note une progression constante vers les valeurs négatives de l'axe 1 et positives de l'axe 2, depuis la classe moins de 25 ans, jusqu'à la classe de 40 à 44 ans ; les 4 classes suivantes sont assez voisines et pourraient être regroupées. On peut interpréter ceci comme un accroissement régulier des revenus jusqu'à 45 ans, puis une stagnation au-delà.

REMARQUE 10 : La dispersion des modalités de la variable CSP sur l'axe 1, nettement plus importante que celle des classes d'âge, indique que la contribution de la CSP à l'inégalité des revenus est beaucoup plus grande que

GRAPHIQUE 3

Représentation graphique des modalités de la CSP et de l'âge selon l'inégalité de la variable revenu (en dimension 2)



Les classes d'âge ont été codées de la façon suivante : *a* : moins de 25 ans ; *b* : de 25 à 29 ans ; *c* : de 30 à 34 ans ; *d* : de 35 à 39 ans ; *e* : de 40 à 44 ans ; *f* : de 45 à 49 ans ; *g* : de 50 à 54 ans ; *h* : de 55 à 59 ans ; *i* : plus de 60 ans.

Rappelons la signification des codes C.S.P. : 0 : agriculteurs exploitants ; 1 : artisans, patrons pêcheurs, petits et gros commerçants, industriels ; 2 : professions non commerciales ; 3 : cadres supérieurs ; 4 : cadres moyens ; 5 : employés ; 6 : ouvriers qualifiés ; 7 : salariés agricoles et ouvriers non qualifiés.

celle de l'âge. Ce résultat peut également être retrouvé par le calcul ; en effet, la variance inter-classes de Z, σ_B^2, peut elle-même être décomposée selon chacune des deux variables X et Y au moyen des formules suivantes :

$$\sigma_B^2 = \frac{1}{n}\sum_{i=1}^{I} n_{i+}(\bar{z}_{i.}-\bar{z})^2 + \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(\bar{z}_{ij}-\bar{z}_{i.})^2$$
$$= \sigma_{B/X}^2 + \sigma_1^2,$$

où $n_{i+} = \sum_{j=1}^{J} n_{ij}$ et $\bar{z}_{i.} = \frac{1}{n_{i+}}\sum_{j=1}^{J} n_{ij}\bar{z}_{ij}$;

$$\sigma_B^2 = \frac{1}{n}\sum_{j=1}^{J} n_{+j}(\bar{z}_{.j}-\bar{z})^2 + \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(\bar{z}_{ij}-\bar{z}_{.j})^2$$
$$= \sigma_{B/Y}^2 + \sigma_2^2,$$

où $n_{+j} = \sum_{i=1}^{I} n_{ij}$ et $\bar{z}_{.j} = \frac{1}{n_{+j}}\sum_{i=1}^{I} n_{ij}\bar{z}_{ij}$.

On déduit des formules analogues pour le carré du coefficient de variation inter-classes τ_B^2, et il est alors clair que celle des deux quantités $\frac{\tau_{B/X}^2}{\tau_B^2}$ ou $\frac{\tau_{B/Y}^2}{\tau_B^2}$ qui est la plus importante correspond à celle des deux variables X ou Y qui contribue le plus, marginalement, à l'inégalité de Z ; en l'occurrence, on obtient $\frac{\tau_{B/X}^2}{\tau_B^2} \simeq 0{,}77$ et $\frac{\tau_{B/Y}^2}{\tau_B^2} \simeq 0{,}11$. Ajoutons que ce résultat est tout à fait conforme à celui obtenu par BOURGUIGNON et MORRISSON [1985] sur des données très comparables (mais pour lesquelles ils disposaient, en plus, des données intra-groupes).

Si l'on s'intéresse maintenant aux interactions entre les deux variables CSP et âge, on constate que les sous-inégalités les plus marquées correspondent à la modalité 7 (et, dans une moindre mesure, aux modalités 0,5 et 6) croisée avec les deux classes d'âge situées en deçà de 30 ans. Quant aux sur-inégalités les plus marquées, elles apparaissent pour la modalité 3 (et, dans une moindre mesure, la modalité 2) croisée avec les classes d'âge situées au-delà de 40 ans. Ces résultats n'ont, bien sûr, rien d'étonnant.

L'intérêt essentiel de la méthode apparaît ainsi clairement : elle fournit une représentation graphique synthétique permettant de mettre en évidence l'interaction des deux variables qualitatives dans la constitution de l'inégalité présentée par la distribution de la variable quantitative.

5 Conclusion

Les méthodes proposées dans cet article ont pour principale originalité d'introduire les techniques de la statistique descriptive multidimensionnelle dans le contexte de l'étude des inégalités liées à plusieurs variables observées simultanément.

L'apport de ces techniques n'est pas suffisamment fondamental pour supplanter les outils classiques que sont les indices synthétiques d'inégalité et l'étude de leurs propriétés, notamment de décomposabilité. Une fois encore, l'approche multidimensionnelle apparaît bien plus complémentaire que concurrente par rapport à l'approche classique (voir notamment BACCINI *et al.* [1987]).

Dans le premier cas considéré (§ 3), le complément essentiel est apporté par la définition des variables d'inégalité principale, qui donnent une vision globale de la situation d'inégalité considérée. Dans le second cas (§ 4), ce sont les graphiques qui apportent ce complément, comme souvent dans les méthodes descriptives.

Les prolongements éventuels de ces méthodes pourraient être, d'une part le remplacement du carré du coefficient de variation par une autre mesure synthétique d'inégalité, en particulier l'indicateur de THEIL, d'autre part la prise en considération de plus de deux variables qualitatives dans le second cas.

ANNEXE 1

Problème de la pondération des individus

La nature du poids $p(\omega)$ que l'on affecte à chaque individu ω dépend du type de données dont on dispose; dans la pratique, trois cas doivent être distingués.

1. *Cas de données brutes*

Chaque individu statistique ω de Ω est une unité statistique élémentaire (use) : individu au sens courant du terme, ménage, exploitation agricole... Dans ce cas, le poids $p(\omega)$ de l'individu ω est choisi en fonction de l'importance qu'on souhaite lui donner (dans la pratique, on utilise par exemple des poids distincts dans les enquêtes par sondage, lorsqu'on réalise *a priori* un échantillonnage à probabilités inégales ou bien lorsqu'on procède *a posteriori* à un redressement d'échantillon); les poids doivent seulement vérifier :

$$\forall \omega \in \Omega,\ p(\omega) > 0 \text{ et } \sum_{\omega \in \Omega} p(\omega) = 1.$$

On a alors

$$\bar{x} = \sum_{\omega \in \Omega} p(\omega)\, X(\omega),\ M_x = n\bar{x}\ (n = \text{card}(\Omega)) \text{ et } q(\omega) = \frac{p(\omega)\, X(\omega)}{\bar{x}}.$$

Dans le cas particulier où $p(\omega) = \frac{1}{n}$, $\forall \omega \in \Omega$, il vient :

$$\bar{x} = \frac{1}{n} \sum_{\omega \in \Omega} X(\omega),\ M_x = \sum_{\omega \in \Omega} X(\omega) \text{ et } q(\omega) = \frac{X(\omega)}{M_x}$$

(définitions données en 2.1).

2. *Cas de données « moyennées »*

Les individus ω de Ω sont dans ce cas des regroupements d'use, en général selon les modalités d'une variable catégorielle (catégorie socio-professionnelle, tranche d'âge, département...); $X(\omega)$ représente la *moyenne* de X au sein de la « catégorie ω » : $X(\omega) = \frac{1}{p(\omega)} \sum_{i \in \omega} p(i)\, X(i)$, où i est une use et $p(i)$ son poids, $p(\omega)$ valant $\sum_{i \in \omega} p(i)$.

On a encore dans ce cas $\bar{x} = \sum_{\omega \in \Omega} p(\omega) X(\omega)$, $M_x = n\bar{x}$ et $q(\omega) = \dfrac{p(\omega) X(\omega)}{\bar{x}}$, où n n'est plus le cardinal de Ω mais le nombre d'use considérées; autrement dit $n = \sum_{\omega \in \Omega} n(\omega)$, avec $n(\omega) = \text{card}(\omega)$.

Dans le cas particulier où $p(i) = \dfrac{1}{n}$, $\forall i \in \omega$, $\forall \omega \in \Omega$, il vient

$$p(\omega) = \frac{n(\omega)}{n}, \quad \bar{x} = \frac{1}{n} \sum_{\omega \in \Omega} n(\omega) X(\omega),$$

$M_x = \sum_{\omega \in \Omega} n(\omega) X(\omega)$ et $q(\omega) = \dfrac{n(\omega) X(\omega)}{M_x}$.

On notera qu'une étude d'inégalité sur des données moyennées entraîne une perte d'information qui peut être importante (il en est d'ailleurs de même pour des données agrégées); notre objet n'est pas ici d'étudier ce problème, et nous renvoyons pour cela à PYATT *et al.* [1980].

3. *Cas de données agrégées*

Comme dans le cas précédent, un individu ω est un regroupement d'use, mais $X(\omega)$ représente maintenant le *cumul* de X au sein de la catégorie ω : $X(\omega) = \sum_{i \in \omega} X(i)$. Ce cumul n'a de sens que si, dans chaque élément ω, les différentes use qui le composent ont le même poids p_ω. Il s'ensuit que $p(\omega) = n(\omega) p_\omega$ (on a toujours $\sum_{\omega \in \Omega} p(\omega) = 1$). La moyenne $\bar{x}$ doit nécessairement s'écrire dans ce cas :

$$\bar{x} = \sum_{i=1}^{n} p(i) X(i) = \sum_{\omega \in \Omega} \sum_{i \in \omega} p_\omega X(i) = \sum_{\omega \in \Omega} p_\omega X(\omega);$$

on pose ensuite $M_x = n\bar{x}$ et $q(\omega) = \dfrac{p_\omega X(\omega)}{\bar{x}}$. Pour réaliser une étude d'inégalité, on doit ici disposer des triplets $\{(n(\omega), p_\omega, X(\omega)); \omega \in \Omega\}$.

L'équipondération des use $\left(p_\omega = \dfrac{1}{n}, \forall \omega \in \Omega\right)$ simplifie sensiblement les choses dans ce cas, puisque l'on obtient :

$$\bar{x} = \frac{1}{n} \sum_{\omega \in \Omega} X(\omega), \qquad M_x = \sum_{\omega \in \Omega} X(\omega) \qquad \text{et} \qquad q(\omega) = \frac{X(\omega)}{M_x}.$$

On retrouve donc, avec l'équipondération, les mêmes formules que dans le cas des données brutes.

On notera que lorsqu'on dispose de données agrégées, il suffit de poser $X'(\omega) = \dfrac{X(\omega)}{n(\omega)}$ pour obtenir les données moyennées correspondantes. Cette transformation est nécessaire pour obtenir $q(\omega) = p(\omega) X'(\omega)$ lorsque la variable X' est de moyenne 1 (voir 3.2).

Décomposition en Valeurs Singulières d'une matrice

Soit T une matrice à n lignes et p colonnes, supposée de rang r. La DVS de T consiste à écrire :

$$T = VS^tU = \sum_{k=1}^{r} s_k v^k {}^t u^k.$$

V (resp. U) est une matrice à n lignes et r colonnes (resp. p lignes et r colonnes), ayant pour colonnes les vecteurs $v^1, \ldots, v^r$ (resp. $u^1, \ldots, u^r$).

$$S = \text{diag}(s_1, \ldots, s_r), \qquad \text{avec} \quad s_1 \geqslant s_2 \geqslant \ldots \geqslant s_r > 0.$$

Les r valeurs propres non nulles de ${}^tTT$ (ou de $T{}^tT$) sont les carrés des valeurs singulières s_k, $k = 1, \ldots, r$.

De plus, $(u^1, \ldots, u^r)$ et $(v^1, \ldots, v^r)$, bases orthonormées « duales » des sous-espaces de $\mathbb{R}^p$ et de $\mathbb{R}^n$ engendrés respectivement par les lignes et les colonnes de T, sont telles que

$${}^tTT\, u^k = s_k^2\, u^k \qquad \text{et} \qquad T{}^tT\, v^k = s_k^2\, v^k,$$

avec $v^k = \frac{1}{s_k} T u^k$ et $u^k = \frac{1}{s_k} {}^tT v^k$.

Si l'on pose $c^k = T u^k = s_k v^k$ et $d^k = {}^tT v^k = s_k u^k$, il vient :

$$T = \sum_{k=1}^{r} \frac{1}{s_k} c^k {}^t d^k.$$

Dans ce contexte, les c^k (resp. les d^k), $k = 1, \ldots, r$, sont appelés composantes principales des lignes de T (resp. des colonnes de T) et vérifient $\langle c^k, c^l \rangle_{\mathbb{R}^n} = \langle d^k, d^l \rangle_{\mathbb{R}^p} = s_k s_l \delta_{kl}$, δ_{kl} désignant le symbole de KRONECKER.

Pour d'autres développements sur la Décomposition en Valeurs Singulières ou sur l'Analyse en Composantes Principales on pourra se reporter à GOWER et DIGBY [1981] ou à VOLLE [1985].

• Références bibliographiques

BACCINI, A., FALGUEROLLES, A. de et QANNARI, E. M. (1986). – « Etude de quelques indices de concentration : un essai de présentation unifiée », *Revue de Statistique Appliquée*, 34, (1), p. 31-44.

BACCINI, A., MATHIEU, J. R. et MONDOT, A. M. (1987). – « Comparaison, sur un exemple, d'analyses des correspondances multiples et de modélisations », *Revue de Statistique Appliquée*, 35, (3), p. 21-34.

BOURGUIGNON, F. (1979). – « Decomposable Income Inequality Measures », *Econometrica*, 47, pp. 901-920.

BOURGUIGNON, F. et MORRISSON, C. (1985). – « Une analyse de décomposition de l'inégalité des revenus individuels en France », *Revue Economique*, 36, p. 741-778.

CANCEILL, G. (1984). – « Les revenus fiscaux des ménages en 1979 », *Economie et Statistique*, 166, p. 39-53.

CAUSSINUS, H. et FALGUEROLLES, A. de (1987). – « Tableaux carrés : modélisation et méthodes factorielles », *Revue de Statistique Appliquée*, 35, (3), p. 35-52.

DAS, T. et PARIKH, A. (1982). – « Décomposition of Inequality Measures and a Comparative Analysis », *Empirical Economics*, 7, pp. 23-48.

DEVILLE, J. C. et MALINVAUD, E. (1983). – « Data Analysis in Official Socio-economic Statistics », *J. R. Statist. Soc.*, series A, 146, pp. 335-361.

DOMENGES, D. et VOLLE, M. (1979). – « Analyse factorielle sphérique : une exploration », *Annales de l'INSEE*, 35, p. 3-84.

ESCOFIER, B. (1984). – « Analyse factorielle en référence à un modèle. Application à l'analyse de tableaux d'échanges », *Revue de Statistique Appliquée*, 32, (4), p. 25-36.

FALGUEROLLES, A. de et HEIJDEN, P. VAN DER (1987). – « Sur l'analyse factorielle des correspondances et quelques-unes de ses variantes », *Revue de Statistique Appliquée*, 35, (3), p. 7-12.

FEI, J., RANIS, G. et KUO, S. (1978). – « Growth and the Family Distribution of Income by Factor Components », *Quaterly Journal of Economics*, 92, pp. 17-53.

GABRIEL, K. R. (1981). – « Biplot Display of Multivariate Matrices for Inspection of Data and Diagnosis », in Interpreting Multivariate Data, Vic Barnett Editor, *Wiley*, New York.

GOWER, J. C. et DIGBY, P. G. (1981). – « Expressing Complex Relationships in Two Dimensions », in Interpreting Multivariate Data, Vic Barnett Editor, *Wiley*, New York.

HEIJDEN, P. VAN DER (1987). – « Correspondence Analysis of Longitudinal Categorical Data », *DSWO Press*, Leiden.

KAKWANI, N. C. (1977). – « Applications of Lorenz Curves in Economic Analysis », *Econometrica*, 45, pp. 719-727.

LEEUW, J. de et HEIJDEN, P. VAN DER (1988). – « Correspondence Analysis of Incomplete Contingency Tables », *Psychometrika*, 53, pp. 223-233.

PYATT, G., CHEN, C. N. et FEI, J. (1980). – « The distribution of Income by Factor Components », *Quaterly Journal of Economics*, 94, pp. 451-474.

QANNARI, E. M. (1983). – « Analyses factorielles de mesures. Applications », *Thèse* de 3e cycle, Université Paul Sabatier, Toulouse.

RAO, V. M. (1969). – « Two Decompositions of Concentration Ratio », *J. R. Statist. Soc.*, series A, 132, pp. 418-425.

SHORROCKS, A. F. (1980). – « The class of Additively Decomposable Inequality Measures », *Econometrica*, 48, pp. 613-625.

SHORROCKS, A. F. (1982). – « Inequality Decomposition by Factor Components », *Econometrica*, 50, pp. 193-211.

SHORROCKS, A. F. (1983). – « The impact of income components on the Distribution of Family Incomes », *Quaterly Journal of Economics*, 97, pp. 311-326.

SHORROCKS, A. F. (1984). – « Inequality Decomposition by Population Subgroups », *Econometrica*, 52, pp. 1369-1385.

THEIL, H. (1967). – « Economics and Information Theory », *North Holland*, Amsterdam.

VOLLE, M. (1985). – « Analyse des données », *Economica*, Paris.