

J. P. BENZÉCRI

Choix des unités et des poids dans un tableau en vue d'une analyse de correspondance

Les cahiers de l'analyse des données, tome 2, n° 3 (1977),
p. 333-352

http://www.numdam.org/item?id=CAD_1977__2_3_333_0

© Les cahiers de l'analyse des données, Dunod, 1977, tous droits réservés.

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

NUMDAM

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

CHOIX DES UNITÉS ET DES POIDS DANS UN TABLEAU EN VUE D'UNE ANALYSE DE CORRESPONDANCE [PONDÉRATION]

par J. P. Benzécri ⁽¹⁾

1. Nous aimons à souligner (cf e.g. [Prat. Corr.] TII A n° 2) que du choix du tableau des données et de leur codage dépend la stabilité d'une analyse. Prétendre perfectionner les calculs statistiques est illusoire si l'on ne s'est d'abord rendu maître de la matière soumise à ces calculs. Et le principal mérite de l'analyse des correspondances est, selon nous, de proposer pour modèle de données le cas d'une correspondance proprement dite, c'est-à-dire d'un tableau de fréquence f_{IJ} rempli par comptage sur une base homogène et exhaustive.

Souvent, c'est le cas notamment en économie, la comparaison par paire des variables, méthode antérieure à l'avènement du calcul électronique, a instauré des pratiques vicieuses de collecte et de codage des informations. Quoique la donnée primaire consiste en bilans qu'il est possible de faire (dans une raisonnable mesure) homogènes et exhaustifs, on ne soumet à l'analyse qu'une liste hétéroclite de rapports (dits ratios) d'après lesquels il n'est pas toujours possible de reconstituer un tableau régulier. Il incombe au statisticien de réformer avec autorité ces mauvais usages. Mais parfois aussi, notamment en biologie, des mesures hétérogènes sont tout ce dont on dispose pour décrire un phénomène important, avec quelque fidélité; un sous-ensemble homogène de mesures laissant au contraire échapper tel ou tel aspect important. Le statisticien doit alors s'efforcer de traiter l'hétérogène.

Aux ensembles de données hétérogènes on peut communément (sans pour autant être assuré de la stabilité de l'analyse) appliquer le codage des variables par rang, ou le codage logique des variables continues par intervalles consécutifs (cf [Prat. Corr.] §§ 1.5 & 1.3). Ce dernier codage autorisé par les travaux de L. Lebart sur les correspondances multiples (cf [Bin. Mult.]), a permis, en conjuguant notamment analyse factorielle et régression, de décrire des interactions très complexes de variables hétérogènes (cf J.P. Bordet, thèse). Le codage étudié dans la présente leçon altère les données moins encore que ne le fait le codage par rang ou le codage logique. Partant d'un tableau de nombres positifs, on se borne à en multiplier chaque colonne (ou aussi chaque ligne) par un coefficient, la suite des coefficients étant déterminée selon certains critères : c'est ce qu'on peut appeler un changement d'unité de mesure, ou de poids (d'importance) pour les variables.

Nous rappelons ci-dessous que le problème des poids se pose même pour une correspondance proprement dite (§ 2) ou un tableau qui paraît homogène (§ 3). Nous proposons ensuite (§ 4) un critère de pondération qui s'applique lorsque les variables (voire les individus) constituent un petit nombre de sous-ensembles homogènes. Après une étude topologique sommaire (§ 5) de l'existence d'un système de poids satisfaisant à ce critère (étude qui n'est pas indispensable au praticien), nous suggérons (§ 6) un algorithme itératif pour déterminer ces poids. Cet

(1) J.P. Benzécri, Professeur de statistique, université Pierre et Marie Curie ; Paris.

algorithme, judicieusement programmé par A. Hamrouni (cf [Pond. Pr.]) a donné satisfaction dans la pratique, et nous en avons déjà publié une application (cf [Myosotis] TI C n° 1, § 1.5) (*). Il n'est cependant pas difficile d'imaginer des situations pathologiques où le système de pondération qu'il serait naturel d'adopter n'est pas une solution limite stable de l'algorithme : de tels contre-exemples sont donnés au paragraphe 7 dans un problème dont la solution est due à P. Cazes.

2. Problème de pondération dans l'étude d'une correspondance proprement dite :

Remarquons d'abord que, même dans le cas d'une correspondance proprement dite, le problème de la pondération peut se poser. Soit I un ensemble de textes, J un ensemble de mots, $k(i, j)$ le nombre de fois que le mot j est employé dans le texte i . Ce nombre $k(i, j)$ est hors de conteste; mais les résultats de l'analyse ne seront indépendants de tout choix arbitraire que pour autant que les ensembles I et J s'imposent comme des tous homogènes à un niveau donné, celui-ci de la littérature, celui-là du lexique. Par exemple si I est l'ensemble des pièces qui se trouvaient au répertoire de la Maison de Molière au début du XIX^e siècle, et J l'ensemble des deux cents mots les plus fréquents dans ce répertoire; ou, mieux encore, si I est l'ensemble de toutes les tragédies de Voltaire, J l'ensemble des adverbess usités dans ces pièces etc... Au contraire, sitôt que l'un ou l'autre des deux ensembles I et J n'est pas un relevé exhaustif, mais résulte d'un choix délibéré, (trente pièces du théâtre classique, vingt mots du vocabulaire de la guerre auxquels tel auteur a consacré sa thèse ...), l'arbitraire s'introduit. Et l'on peut dire qu'il s'agit ici d'un système de poids, pour autant que choisir un ensemble I , c'est affecter de poids l'ensemble I' , plus vaste que I , des éléments i possibles a priori.

En effet, représentons-nous I' comme un nuage dans quelque espace topologique I'' : I'' sera par exemple l'ensemble *potentiel* de toutes les pièces qu'on eût pu écrire sans trop s'écarter des canons classiques; I' l'ensemble *actuel* des pièces effectivement écrites (et dont le texte s'est conservé); I l'ensemble *choisi*. Dire que I' est représentatif de I'' , revient à dire qu'au cours de l'âge classique, nos poètes ont à peu près épuisé les possibilités du genre dramatique tel qu'ils le concevaient. C'est là une hypothèse qui ne s'impose nullement: entre Polyeucte, Andromaque et l'Orphelin de la Chine, trop grandes sont les différences pour que l'on puisse croire le genre seulement exploré; on douterait plutôt que malgré leur rigueur, les règles classiques suffisent à définir I'' : nous l'avons dit ailleurs, seul un genre lapidaire, abondamment pratiqué, tels les dystiques extrême-orientaux, offre peut-être l'exemple d'un potentiel I'' exactement défini et bien représenté par l'actuel I' .

Acceptons cependant l'hypothèse que I' représente I'' en ce sens topologique, qu'assez près de tout point de I'' se trouve un point de I' . Si l'ensemble I' est d'un cardinal trop élevé pour servir lui-même de base à un tableau de correspondance, on doit faire choix d'un ensemble I qui représente I' , au sens où nous venons de supposer que I' représente I'' ; ainsi I représentera I'' , avec seulement moins de précision que ne le fait I' . Mais ici se pose un problème de pondération.

Supposons en effet que l'on double le nombre des éléments d'une certaine région R' de I' , acceptés dans l'ensemble I , base du tableau $I \times J$ qu'on choisit d'analyser. Cela équivaut, à peu près, à ne pas modifier I , mais à multiplier par un coefficient 2 les nombres $k(i, j)$,

(*) Le programme [Pond. Pr] paraît dans le présent cahier : cf pp. 353-359; quant aux applications, cf *Erodium* §§ 1.2 & 2, Vol II n° 1, pp 98-103.

tels que $i \in R'$. Pour que les résultats de nos calculs soient univoquement définis (avec seulement une précision plus ou moins grande selon le volume d'information traité), il ne suffit pas que I'' (ou J'' ensemble potentiel d'où provient J) soit un espace topologique assez bien défini, il faut que cet espace soit muni d'une loi de probabilité $p_{I''}$, (autrement dit, il faut que I'' soit un espace probabilisé) : I' (ou, après lui, I) sera un représentant de I'' , s'il est un échantillon suffisamment dense de la loi $p_{I''}$. Ainsi, non seulement nous sommes conduits à assujettir l'ensemble I à des conditions plus restrictives, mais encore nous devons supposer qu'au niveau du potentiel sont satisfaites des hypothèses plus riches, sinon véritablement plus restrictives. Précisons ce point.

Nous doutions que les règles classiques et, avec elles, l'usage classique, définissent exactement un ensemble potentiel, I'' , de pièces classiques. N'est-il pas plus hasardeux encore de parler d'un espace probabilisé ? Ce n'est pas sûr. Mathématiquement un espace probabilisé est plus qu'un espace, car c'est un espace avec une mesure. Mais dans la nature, chercher un espace probabilisé a ceci d'avantageux qu'on est dispensé de définir exactement la frontière de I'' - entre ce qui est pièce et ce qui ne l'est pas - Quant à la mesure $p_{I''}$, on peut se contenter d'une notion assez vague de probabilité. En ce sens, le cadre ensembliste et logique est moins propre à recevoir la nature que ne l'est le cadre probabiliste, pourvu que celui-ci soit convenablement assoupli (condition certes essentielle, cf [Principes] T II A n° 1 § 1°) : ceci nous semble manifeste, notamment, en linguistique (cf Linguistique et Mathématique : le siège des lois du langage).

Quoiqu'il en soit des hypothèses, il reste qu'il n'est pas facile de choisir I : car ce choix suppose une bonne connaissance sinon de I'' , au moins de I' or c'est justement parce qu'on connaît mal I' qu'on entreprend l'analyse, et c'est principalement pour éviter d'examiner I' avec trop de détails qu'on se restreint à I .

3. L'analyse des tableaux homogènes :

Quittons maintenant le cas des correspondances proprement dites pour considérer d'autres tableaux : d'abord des tableaux homogènes, puis des tableaux hétérogènes. Souvent un tableau homogène peut se ramener à une correspondance; voici quelques exemples. Soit I un ensemble dont nous appellerons les éléments des individus; J un ensemble de qualités; $\{k(i, j)\}$ une matrice logique ou tableau de description : $k(i, j) = 1$ si l'individu i possède la qualité j ; et $k(i, j) = 0$, sinon. Le choix des ensembles I et J pose un problème que nous venons d'examiner dans le cas des correspondances avec, pour l'ensemble des qualités, la difficulté suivante signalée par M.S. Watanabe (in *Information and Prediction in Science*, edited by Dockx and Bernays, 1965).

Plaçons-nous dans un cadre purement logique et supposons qu'il existe exactement N descriptions possibles pour les éléments de la classe qui nous intéresse (ceci équivaut à identifier actuel et potentiel, à circonscrire indûment le possible : c'est là une infirmité inhérente au point de vue logique). On peut dire alors qu'il y a 2^N prédicats, autant que de parties de l'ensemble des descriptions : chaque élément possède exactement 2^{N-1} de ces prédicats (autant qu'il y a de parties contenant sa description); et deux éléments dont les descriptions ne sont pas identiques divisent en quatre parties égales (oui-oui, oui-non, non-oui, non-non...) l'ensemble des prédicats. Dans ce cadre, il n'y a aucune notion de proximité possible. Pour distinguer et unir, il faut donc considérer la classe des prédicats non comme l'ensemble Part(Desc) des parties de l'ensemble Desc des N descriptions possibles, mais comme un ensemble naturel J de propriétés qui, d'une certaine manière, tombent sous le sens.

Rappelons ici que si l'ensemble J est muni d'une partition G , i.e. si :

$$G \subset \text{Part}(J); J = \cup \{g | g \in G\};$$

$$\forall g, g' \in G : (g \cap g' = \emptyset) \Leftrightarrow (g \neq g'),$$

la matrice de description logique $k(i, j)$ définit une correspondance sur $I \times G$:

$$k(i, g) = \sum \{k(i, j) | j \in g\}.$$

Et c'est croyons-nous faire choix d'un excellent critère de classification sur J , que de demander que, (le cardinal de G , ou nombre des classes, étant fixé) soit maxima la distance entre le tableau $k(i, g)$ et l'indépendance (distance lien₂(I, G) = $\| \bar{p}_{IG} - p_{IPG} \|^2_{p_{IPG}}$, cf [Inf. Tab.] TI B n° 5 § 2.1.1) : ainsi la même distance du χ^2 sert à transformer la matrice logique k_{IJ} en tableau de correspondance k_{IG} puis à analyser celui-ci.

Mais il importe de distinguer ici entre distance du χ^2 , et épreuve du χ^2 . Lors même qu'un tableau homogène se laisse naturellement interpréter comme un tableau de correspondance, les $k(i, j)$ étant tous entiers, ces entiers ne résultent pas du comptage d'événements aléatoires indépendants; chaque unité a un poids plus grand que dans le cas d'une correspondance proprement dite. Dire qu'un sujet résout trois des questions d'un problème, qu'un terrain contient trois des espèces d'un genre, c'est beaucoup plus que de dire que tel événement aléatoire s'est produit trois fois. L'épreuve du χ^2 est donc de peu de secours pour décider du nombre des facteurs significatifs. Souvent d'ailleurs la théorie de la décision statistique demande trop d'hypothèses pour être capable de fournir plus que des indices (cf [Epr. Val.] TII B n° 8 § 1 *in fine*).

Voici un autre type de tableau homogène : comme ci-dessus, I est appelé ensemble des individus; J est un ensemble de mesures de même nature; $k(i, j)$ est le résultat obtenu en prenant la mesure j sur l'individu i . Le choix des ensembles I et J ne nous pose pas de problèmes nouveaux. Le choix de l'unité n'a pas d'importance, puisque les mesures étant de même nature elle est la même pour toutes, et que l'on ne change pas les résultats de l'analyse si on multiplie simultanément tous les $k(i, j)$ par un même nombre (changement d'unité). Il faut cependant préciser ce qu'on entend par mesures de même nature. Soit à décrire par des mesures de longueur, un animal : la longueur de la patte, le périmètre du crâne, le diamètre du genou... doivent assurément être mesurés dans la même unité (cf TI C n° 5, 6, 7, 8). Mais que dire du diamètre d'un globule rouge, ou de celui du cylindre axe d'un neurone ? sans cesser d'être des longueurs, ce ne sont pas des mesures de même nature que celles considérées d'abord. La même difficulté se présente si l'on mesure dans un milieu donné i - e.g. le sang d'un patient - les taux de divers produits j . A des concentrations très différentes en j et j' peuvent correspondre des effets équivalents : or ce sont les effets qui nous intéressent, non les concentrations, les masses elles-mêmes. Mais celles-ci étant seules mesurables, on doit recourir à des expédients pour que le tableau analysé soit une image, à peu près fidèle des effets. Par exemple, on choisira pour unité de mesure du taux de j la somme du seuil de précision P_j avec lequel on fait cette mesure, et du seuil de signification S_j que le spécialiste (ici un médecin) déclare utiliser dans son appréciation empirique des mesures : autrement dit, à $k(i, j)$ on substituera $k(i, j)/(P_j + S_j)$.

Il est intéressant de citer encore un exemple, dont nous empruntons la discussion à G. Thauront.

"Dans une enquête en vue d'une typologie des communes françaises par rapport à leur équipement, nous avons deux sortes de données mesurées en mètres :

- 1°) des longueurs d'équipement (ex : longueur de voiries)
- 2°) des distances à un équipement ponctuel : (ex : distance au plus proche hôpital).

L'hétérogénéité est évidente. D'abord les concepts de "longueur" et "distance" sont distincts. Ensuite, une augmentation de la longueur de voiries est une amélioration alors que c'est une diminution de la distance au plus proche hôpital qui est une amélioration. Ce deuxième argument n'est donné que pour mémoire car s'il est valable ici, il ne l'est pas en général, (exemple : des recettes et des dépenses peuvent être considérées comme homogènes) et de plus nous avons appris à nous défier des orientations a priori entre bien et mal (cf enquête [Colombie] TII C n° 6). D'où un critère nécessaire d'homogénéité : il faut que lors de la réunion de deux individus proches, si cela a un sens, (e.g. agrégation en une commune unique de deux communes limitrophes) le calcul des mesures sur l'individu réunion se fasse à partir des mesures sur les individus réunis, suivant la même loi.

Ici, ce n'est pas le cas : les longueurs de voiries s'ajoutent alors que la distance au plus proche hôpital est une moyenne des distances".

4. Critères de pondération dans l'analyse des tableaux hétérogènes :

Venons-en maintenant aux données proprement hétérogènes : par exemple aux tableaux dont certaines colonnes contiennent des mesures de quantités, tandis que d'autres contiennent une réponse logique (oui, non; 1,0) ou des modalités de qualité (modalités qu'il convient, suivant des principes que nous avons considérés parmi ceux de la taxinomie, de convertir en réponses logiques simples...). Tout scrupule méthodologique mis à part, l'empirisme est embarrassé pour choisir des coefficients de pondération. Avant de tenter de venir à son secours, formulons une fois de plus, nos réserves devant ces tableaux hétérogènes. Mêler des données trop diverses pour être représentatives (Cf supra) de quelque ensemble potentiel que ce soit, c'est fondre en un tout les observations avec les hypothèses : il n'est pas sûr qu'aucun algorithme statistique suffise à épurer ce conglomerat. Dans les résultats de nos calculs le spécialiste qui a fourni les données risque de ne retrouver que sa propre image; tout au plus, en constatant qu'un statisticien non prévenu armé d'un algorithme universel a déchiffré sur un tableau de nombres les lignes de sa doctrine, acquerrera-t-il dans notre science une confiance qu'il nous incombera ensuite de ne pas décevoir ! Tandis que l'étude, aussi exhaustive que possible d'un domaine homogène, si peu séduisant soit-il a priori, peut conduire à la découverte de dimensions vraiment nouvelles.

Ces réserves faites, considérons un tableau $\{k(i, j)\}$ de nombres réels positifs, indicé par le produit de deux ensembles finis I et J sur lesquels nous formulons l'hypothèse d'homogénéité partielle la plus générale : I (resp J) est muni d'une partition Q (resp G) en classes homogènes (cf. supra); on note $q(i)$ resp $g(j)$ l'unique classe $q \in Q$, (resp $g \in G$), telle que $i \in q$ (resp $j \in g$). Cette situation se présente par exemple si I est un ensemble de patients formant des classes distinctes (e.g. des nourrissons, des hommes adultes, des femmes adultes, etc...) et J un ensemble de variables physiologiques susceptibles elles aussi d'être classées... Du point de vue mathématique rien n'interdit de considérer les cas extrêmes : $Q = \{I\}$ (homogénéité de I) et $Q = \{\{i\} | i \in I\}$ (hétérogénéité a priori, maxima).

Faire un changement d'échelle (ou de poids, ou d'unité) sur chacun des ensembles I et J , c'est faire un choix de deux systèmes $\{\lambda(i) | i \in I\}$ et $\{\mu(j) | j \in J\}$ de nombres strictement positifs, et substituer au tableau initial le tableau des $k(i, j) \times \lambda(i) \times \mu(j)$. Du fait de l'hypothèse d'homogénéité partielle, il est naturel de ne pas modifier les pondérations relatives des éléments d'une même classe homogène. Autrement dit, il suffit de chercher deux systèmes $\{\lambda(q) | q \in Q\}$ et $\{\mu(g) | g \in G\}$, et de poser : $\lambda(i) = \lambda(q(i))$; $\mu(j) = \mu(g(j))$. Pour déterminer ces deux systèmes, il faut avoir des exigences a priori sur l'importance relative que l'on veut donner dans l'analyse à chacune des classes homogènes; traduire ces exigences en équations reliant les $\lambda(q)$ et les $\mu(g)$; puis résoudre (au moins approximativement) ces équations.

On sait (cf [Prat. Cor.] § 2.5) que la trace $\|p_{IJ} - p_I p_J\|_{p_I p_J}^2$ qui mesure l'écart dont l'analyse factorielle prétend rendre compte, peut s'exprimer comme une somme indicée soit par I soit par J . On a donc défini la contribution absolue $c(j)$ de l'élément j de J à la trace par la formule :

$$c(j) = p_j \rho^2(j) = p_j \|p_I^j - p_I\|_{p_I}^2 \\ = k(j) \sum \{ [(k(i, j)/k(j)) - (k(i)/k)]^2 / k(i) | i \in I \},$$

(i.e. c'est le produit par la masse p_j du carré de la distance, au sens du χ^2 , entre le profil p_I^j et le profil p_I). On définit de même la contribution $c(i)$ de l'élément i . L'on a bien :

$$\text{Trace} = \sum \{c(j) | j \in J\} = \sum \{c(i) | i \in I\}.$$

Et il est naturel de définir la contribution d'une classe comme la somme des contributions de ses éléments :

$$c(g) = \sum \{c(j) | j \in g\}; \quad c(q) = \sum \{c(i) | i \in q\}.$$

On est ainsi conduit au problème suivant : déterminer les poids $\lambda(i)$ et $\mu(j)$ de telle sorte que, dans la trace du tableau $\{k(i, j) \lambda(i) \mu(j)\}$, les contributions $\{c(q) | q \in Q\}$, $\{c(g) | g \in G\}$ soient proportionnelles à des systèmes de nombres réels positifs donnés a priori. Nous ne prétendons pas résoudre le système d'équations qui exprime cette proportionnalité; mais seulement suggérer aux théoriciens une voie pour discuter l'existence et l'ordre de multiplicité des solutions (§ 5); puis proposer une formule approchée (§ 6) qui, malgré son intérêt pratique, n'est pas d'une application universelle (§ 7).

5. Esquisse d'une étude topologique (cette étude n'est pas indispensable au praticien) :

Notons λ_Q (resp μ_G , c_Q , c_G), le système des nombres $\lambda(q)$ (resp $\mu(g)$, $c(q)$, $c(g)$) réduits à avoir pour somme 1; et notons P_S (cf TII B n° 4 § 1) le simplexe ensemble des lois de probabilité (ou systèmes de masses positives de somme 1) sur l'ensemble S (ici $S = Q$ ou G). Notre objet est d'étudier l'application C du produit $P_Q \times P_G$ dans lui-même.

$$(\lambda_Q, \mu_G) \rightarrow (c_Q, c_G),$$

(à un couple de systèmes de poids, C fait correspondre le couple des systèmes de contributions de classe, réduit à avoir pour somme 1). Pour la discussion d'existence et de multiplicité, on va, selon l'usage, recourir à des méthodes topologiques. On va montrer que l'application C a en général pour indice 1, d'où il résultera qu'il existe au moins un

couple (λ, μ) ayant une image $C(\lambda, \mu)$ donnée (il y aura plusieurs couples si, le jacobien de C changeant de signe, certaines solutions doivent être comptées avec un indice négatif; nous verrons que cette irrégularité se produit : cf § 7, 4°, 6°, 8°).

Supposons d'abord que chacune des classes q ou g compte plus d'un élément. On est assuré ainsi que $c(g)$ (ou $c(q)$) ne s'annule que si $\lambda(g)$ (resp $\lambda(q)$) s'annule, et que, au voisinage de $\lambda(g) = 0$, $c(g)$ est un infiniment petit comparable à $\lambda(g)$ (le rapport de ces deux infiniment petits reste borné). Sous cette hypothèse, l'application C envoie dans elle-même chaque face du polyèdre $P_Q \times P_G$: elle met en correspondance biunivoque les sommets; donc sur chaque arête et plus généralement sur toute face et en particulier sur l'intérieur du polyèdre elle a pour indice 1.

La situation se complique quand une classe, e.g. $g^\circ = \{j^\circ\}$, est réduite à un élément : car alors, si $\mu_{g^\circ} = 1$ (les autres μ_g étant nuls : μ_g est au sommet de g° du simplexe P_G) le profil moyen p_I coïncide avec le profil $p_I^{j^\circ}$. Le terme $c(g^\circ)$ s'annule comme un infiniment petit du second ordre en $(1 - \mu_{g^\circ})$, tandis que les autres termes $c(g)$ s'annulent (sauf si $g \neq g^\circ$ ne contient que des j dont le profil se trouve coïncider avec celui de j° ...) comme un infiniment petit du premier ordre en μ_g . Quand μ_g coïncide avec le sommet g° , C_G bien loin d'être en ce même sommet est sur la face opposée du simplexe P_G .

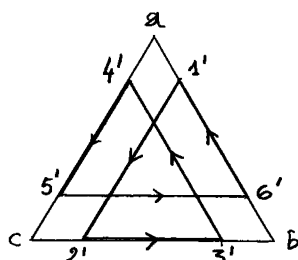
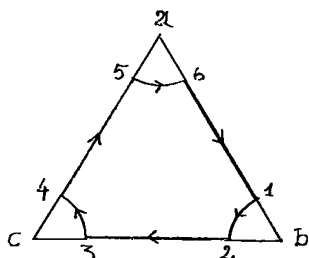
Prenons un exemple : $Q = \{I\}$, (I est homogène); $J = \{a, b, c\}$; $G = \{\{j\} | j \in J\}$, (J est totalement hétérogène; et toutes les classes g sont singulières). Le simplexe P_Q étant réduit à un point, on peut dire qu'on étudie une application D du triangle $P_G = (abc)$, dans lui-même. Les trois profils p_I^j sont bien déterminés (indépendants de μ_G); les probabilités p_j sont liées aux $\mu(j)$ par la formule :

$$p_j = \mu(j) \sum \{k(i, j) | i \in I\} / \sum \{k(i, j) \mu(j) | i \in I, j \in J\}.$$

On prendra les $\{p_j\}$ comme système de coordonnées homogènes dans P_G , source de l'application C ; et on étudiera l'application du triangle P_G dans lui-même qui en coordonnées homogènes s'écrit :

$$P_J = \{p_a, p_b, p_c\} \rightarrow \{p_j | p_I^j - p_I\|^2 | j \in J\} = \{c(a), c(b), c(c)\}.$$

On va montrer que cette application a pour indice (-2) . Comme il faut éviter que p_J ne soit à un sommet du triangle $(a b c)$, considérons le contour 1 2 3 4 5 6 1, et montrons que son image a la forme du contour 1' 2' 3' 4' 5' 6' 1', que nous avons figuré : puisque ce dernier contour fait deux fois le tour du triangle dans le sens opposé à $a b c$, on aura bien montré que l'application C a pour indice (-2) .



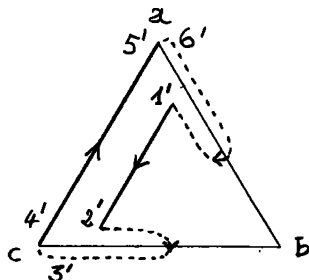
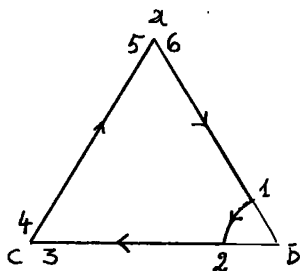
Au point 1, on a $p_c = 0$; $p_b \approx 1$; $p_a = \epsilon$; parce que $p_c = 0$, $c(c) = 0$; parce que p_b ne diffère de 1 que par un infiniment petit du premier ordre, la distance entre p_I et p_I^b est un infiniment petit du deuxième ordre $O(\epsilon^2)$; enfin $c(a)$ est, comme p_a , un infiniment petit du premier ordre; d'où la position figurée de $1'$. Quand on passe de 1 à 2, $c(b)$ reste un infiniment petit du deuxième ordre $O((p_a + p_c)^2)$; tandis que l'on a :

$$c(a) \approx p_a \|p_I^a - p_I^b\|^2; \quad c(c) \approx p_c \|p_I^c - p_I^b\|^2,$$

(où les normes sont calculées dans p_I , muni de la métrique du χ^2 de centre $p_I \approx p_I^b$). Quand on passe de 2 à 3, p_a , donc $c(a)$ est identiquement nul; et ainsi de suite.

La même méthode s'étend au cas où $Q = \{I\}$, $G = \{\{j\} | j \in J\}$ avec J de cardinal quelconque : on démontre par récurrence que l'indice de l'application $p_J \rightarrow c_J$ est $(\text{Card}(J) - 1)$. Cependant notre calcul peut tomber en défaut si dans p_I certains des p_I^j sont confondus.

Dans certains cas, l'indice de C peut être nul : c'est ce qui se produit, e.g., si $Q = \{I\}$; $G = \{a, b, c\}$, la classe b étant seule réduite à un élément; sur le contour, on peut en procédant comme ci-dessus, montrer que l'application C a l'allure indiquée sur la figure :



L'application C laisse fixes les sommets a et c , et replie, en quelque sorte, l'angle b .

En général notons z_Q et z_G le nombre des classes de Q et G -réduites

à un seul profil. Sauf dégénérescences supplémentaires que nous n'envisagerons pas (coïncidence entre les profils de deux classes...), le bord de $P_Q \times P_G$ s'envoie dans lui-même par C et le degré de l'application C est $(1 - z_Q)(1 - z_G)$. Pour le voir il suffit de calculer l'indice au voisinage d'un sommet (δ_Q^q, δ_G^g) de $P_Q \times P_G$: pour que $C(\lambda, \mu)$ tende vers ce sommet, il faut que (λ, μ) tende lui aussi vers un sommet $(\delta_Q^{q'}, \delta_G^{g'})$ (en bref : q', g'), soit que $q' = q$ (ou $g' = g$) si q est une classe non triviale; soit que q' soit une classe triviale distincte de q , et que (λ, μ) tende vers q' suivant la direction qq' ; l'indice étant 1 si q' et g' sont tous deux de même nature (tous deux triviaux, ou tous deux non triviaux...) et -1 sinon.

6. Une formule approchée de changement d'échelle :

Quittons cet exercice de topologie, pour proposer une méthode de calcul. Considérons comme ci-dessus un tableau $\{k(i, j)\}$ de nombres réels indicé par le produit de deux ensembles finis I et J , chacun muni d'une partition notée respectivement Q et G . Supposons donnés *a priori* deux systèmes de nombres positifs de somme 1, $d_Q = \{d_q | q \in Q\}$ et $d_G = \{d_g | g \in G\}$, proportionnels aux contributions que nous désirons voir apporter dans la trace par les classes q et g . Notons, comme précédemment, $c(q)$, ou $c(g)$, la contribution effective d'une classe calculée d'après le tableau $\{k(i, j)\}$ non modifié (considéré avant tout changement d'unités). Nous suggérons le changement d'échelle suivant :

$$\lambda_q = d_q / c(q); \quad \mu_g = d_g / c(g);$$

et, (en utilisant le signe $:=$ pour indiquer l'affectation au membre de gauche d'une nouvelle valeur exprimée à droite en fonction des valeurs antérieures), le tableau à analyser des valeurs modifiées s'écrit :

$$k(i, j) := k(i, j) \times d_{q(i)} \times d_{g(j)} / (c(q(i)) \times c(g(j))).$$

Cette formule conduit exactement au système de contributions relatives demandé si le changement d'unité ne modifie que les masses relatives p_i , p_j mais laisse invariantes les distances $\|p_i^j - p_i^i\|^2, \|p_j^i - p_j^j\|^2$. Pour examiner cette situation, plaçons-nous dans un cas tout à fait particulier. Supposons que I et J sont des ensembles produits : $I = Q \times T$, $J = G \times S$; et que l'on a un tableau k sur $T \times S$ et deux fonctions (fonctions de poids) l et m sur Q et G telles que :

$$\forall i = (q, t) \in I, \forall j = (g, s) \in J : k(i, j) = k(t, s) l(q) m(g).$$

Cette condition exprime l'hypothèse que les classes q (ainsi que les classes g) se correspondent entre elles biunivoquement; à chaque élément $(q, t) \in I$ correspond un élément $(q', t) \in I$, qui a même profil que lui, et aussi même masse relative dans sa classe :

$$k(q, t) / k(q', t) = k(q, t') / k(q', t') = l(q) / l(q').$$

Mais pour les conséquences qui nous intéressent, il suffit de supposer que $\{k_i^j | i \in q\}$ et $\{k_j^i | i \in q'\}$, (k_i^j muni de la masse k_i), constituent dans P_J des nuages de points à peu près équivalents : des nuages qui peuvent être regardés comme des échantillons d'une même loi. Et c'est pourquoi le cas extrême où nous sommes placés peut apparaître comme la formalisation d'un cas réel. La forme particulière du tableau n'est pas altérée par un changement de poids $\{\lambda(q)\}$, $\{\mu(g)\}$: bien plus, en posant $\lambda = 1/l$, et $\mu = 1/m$, il vient $k(i, j) = k(t, s)$. La distance distributionnelle entre deux éléments de I (resp J) ne dépend que de

leur deuxième coordonnée dans T (resp S). Le centre de gravité p_J (ou p_I) de l'ensemble des profils est invariant par changement de poids $\lambda(q)$, $\mu(g)$. Les $c(q)$ (resp $c(g)$) sont, après changement d'échelle, proportionnels aux masses $\lambda(q) \times l(q)$, (resp $\mu(g) \times m(g)$). Notre formule de changement d'échelle est ici rigoureusement exacte.

Dans un cas réel, la formule de changement de poids ne conduira pas exactement aux systèmes de contributions relatives désirées: si l'approximation obtenue n'est pas satisfaisante, on pourra refaire un changement d'échelle suivant la même formule à partir des nouvelles valeurs de $k(i,j)$.

Dans la pratique, cet algorithme itératif de changement d'échelle a été exclusivement appliqué au cas où l'ensemble des individus (noté ici J) ne forme qu'une seule classe homogène (la partition $G = \{J\}$ ne comprend qu'une seule classe); et les poids relatifs d_q attribués aux diverses classes q de variables sont choisis suivant l'une des deux règles suivantes :

1°) toutes les classes sont *a priori* d'égale importance :

$$\forall q \in Q : d_q = 1/\text{Card } Q.$$

2°) On attribue à la classe q une contribution dans l'analyse globale proportionnelle à la trace du tableau partiel k_{qxJ} . Cette dernière règle sera suivie si les classes q sont d'effectifs très inégaux; ou encore lorsque certains des tableaux partiels k_{qxJ} semblent apporter une information très faible dont la trace de k_{qxJ} pourrait (faute d'une meilleure évaluation) donner un ordre de grandeur.

Le programme écrit par A. Hamrouni offre à l'utilisateur le choix entre ces deux règles (cf [Pond. Pr.] pp 353-359).

7. Stabilité de l'algorithme de changement d'échelle :

Un assez grand nombre d'applications diverses (à des cas où $\text{Card } G = 2$ ou 3) nous ont convaincu de l'intérêt que présente le programme de A. Hamrouni : toujours l'algorithme de pondération a convergé, en une dizaine d'itérations au plus, vers une solution satisfaisante. Cependant il est en théorie possible qu'il n'en soit pas ainsi. L'étude détaillée de la convergence est toutefois rendue difficile parce que, en changeant les poids relatifs p_i des caractères i de I , on change aussi la loi p_J , donc la métrique dont est muni l'espace R_J où est le nuage des p_J^i . C'est pourquoi, dans le problème ci-dessous, on se place principalement dans un espace euclidien E muni d'une métrique fixée; et l'on considère dans E un ensemble hétérogène I , réunion de deux sous-ensembles homogènes I_1 et I_2 . On cherche alors à donner au rapport des masses de ces deux sous-ensembles une valeur telle que I_1 et I_2 apportent à l'inertie totale du nuage des contributions égales (ou plus généralement ayant entre elles un rapport prescrit). Dans la 8° question on montre brièvement que les situations typiques étudiées ici, se rencontrent aussi en analyse des correspondances, alors que l'espace $E = R_J$

est muni d'une métrique euclidienne qui varie avec les pondérations relatives de I_1 et I_2 .

7.1. Enoncé du 2° problème proposé au D.E.A. de Statistique en Juin 1972.

Soit E un espace euclidien (espace vectoriel muni d'une forme quadratique de distance); I_1, I_2 des parties finies de E . On suppose donné sur I_1 une fonction positive m_1 assignant à chaque point $i \in I_1$ une masse, en sorte que la masse totale soit 1 :

$$\forall i \in I_1 : m_1(i) \geq 0; \sum \{m_1(i) | i \in I_1\} = 1;$$

De même on suppose donné sur I_2 une fonction m_2 telle que :

$$\forall i \in I_2 : m_2(i) \geq 0; \sum \{m_2(i) | i \in I_2\} = 1.$$

On note g_1 le centre de gravité du nuage N_1 des points $i \in I_1$ affectés des masses $m_1(i)$; et S_1 le moment total d'inertie de ce nuage par rapport à son centre :

$$g_1 = \sum \{m_1(i) i | i \in I_1\}$$

$$S_1 = \sum \{m_1(i) \|g_1 i\|^2 | i \in I_1\}$$

On définit de même g_2 et S_2 pour le nuage N_2 des points de I_2 affectés des masses m_2 . On suppose désormais S_1 et S_2 non-nuls.

Enfin étant donné deux nombres positifs μ_1 et μ_2 de somme 1 :

$$0 \leq \mu_1; 0 \leq \mu_2; \mu_1 + \mu_2 = 1 \text{ (i.e. } \mu_2 = 1 - \mu_1),$$

on définit un nuage global N formé des points de l'ensemble $I = I_1 \cup I_2$ affecté des masses $m(i)$ données par la formule suivante :

$$\forall i \in I_1 : m(i) = \mu_1 m_1(i)$$

$$\forall i \in I_2 : m(i) = \mu_2 m_2(i) = (1 - \mu_1) m_2(i)$$

1°) Calculer en fonction de S_1, S_2, μ_1 et de la distance $R = g_1 g_2$, le moment d'inertie total $S(\mu_1)$ du nuage N par rapport à son centre.

2°) Quelles sont les contributions respectives $C_1(\mu_1)$ et $C_2(\mu_1)$ des sous-ensembles I_1 et I_2 à cette inertie totale $S(\mu_1)$.

3°) Soit ρ un nombre positif. Est-il possible de choisir μ_1 et μ_2 ($\mu_2 = 1 - \mu_1$) de telle sorte que :

$$C_1(\mu_1)/C_2(\mu_1) = \rho ?$$

Ce problème peut-il admettre plusieurs solutions ?

4°) On suppose ici que l'espace E est la droite réelle (munie de la distance usuelle); on pose :

$$I_1 = \{(-1-a), (-1+a)\}; m_1(-1-a) = m_1(-1+a) = 1/2$$

$$I_2 = \{(1-a), (1+a)\}; m_2(1-a) = m_2(1+a) = 1/2$$

(i.e. N_1 est formé de deux points de masse égale à $1/2$, séparés par la distance $2a$, et situés de part et d'autre du point d'abscisse -1). Appliquer les formules trouvées en 1° et 2°; et dire, en fonction de a , pour combien de valeurs de μ_1 il y a égalité des contributions ($\rho = 1$).

5°) Afin de résoudre itérativement l'équation $C_1(\mu_1)/C_2(\mu_1) = \rho$, on part d'une valeur initiale quelconque $\mu_1(0)$; et on calcule successivement $\mu_1(1)$, $\mu_1(2)$ etc... comme il est expliqué ci-dessous (en notant $\mu = \mu_1(n-1)$) :

$$\mu_1(n)/(1 - \mu_1(n)) = (C_1(\mu)/\rho C_2(\mu))^\alpha (\mu/(1 - \mu))$$

Choisir dans cette formule l'exposant α pour que, dans le cas particulier où $R = 0$ (sous-nuages N_1 et N_2 de même centre), la valeur $\mu_1(1)$ soit bonne quelle que soit la valeur $\mu_1(0)$ prise au départ.

6°) On revient maintenant au cas particulier considéré en 4°, pour étudier l'algorithme itératif proposé en 5°; et on pose $\rho = 1$. Soit $\mu_1(0)$ voisin de $1/2$ (e.g. : $\mu_1(0) = (1/2) - \epsilon$). Discuter selon la valeur de a si $\mu_1(1)$ sera plus proche ou plus éloigné de $(1/2)$ que ne l'est $\mu_1(0)$. Dans le cas où l'équation $C_1(\mu_1)/C_2(\mu_1) = 1$ admet une ou plusieurs autres solutions que $(\mu_1 = 1/2)$, discuter de même la valeur de $\mu_1(1)$ si $\mu_1(0)$ est proche d'une de ces solutions. Que peut-on dire de la stabilité de l'algorithme de 5° ?

7°) Reprendre la discussion de 6° dans le cas général (ρ nombre positif quelconque; et N_1 , N_2 , nuages quelconques).

8°) Soit J un ensemble fini; k_1 et k_2 des tableaux de nombres positifs sur $I_1 \times J$ et $I_2 \times J$ respectivement; on suppose que :

$$\forall i \in I_1 : k_1(i) = \sum \{k_1(i, j) | j \in J\} = m_1(i)$$

$$\forall i \in I_2 : k_2(i) = \sum \{k_2(i, j) | j \in J\} = m_2(i).$$

On note, comme précédemment, $I = I_1 \cup I_2$; et pour $\mu_1 \in (0, 1)$ on définit un tableau k_{IJ} :

$$\forall i \in I_1, \forall j \in J : k(i, j) = \mu_1 k_1(i, j)$$

$$\forall i \in I_2, \forall j \in J : k(i, j) = (1 - \mu_1) k_2(i, j).$$

On note $Q_1(\mu_1)$ et $Q_2(\mu_1)$ les contributions respectives des sous-ensembles I_1 et I_2 à l'inertie totale du nuage I associé au tableau de correspondance k_{IJ} . Dans quelle mesure les considérations des questions 1° à 7° s'appliquent-elles à la recherche d'un nombre μ_1 tel que $Q_1(\mu_1)/Q_2(\mu_1)$ ait une valeur ρ donnée a priori ?

7.2. Correction de la seconde épreuve D.E.A. de Statistique Juin 1972.

7.2.1. Le centre de gravité g du nuage N s'écrit :

$$\begin{aligned}
 g &= \Sigma \{m(i)i \mid i \in I\} \\
 &= \mu_1 \Sigma \{m_1(i)i \mid i \in I_1\} + \mu_2 \Sigma \{m_2(i)i \mid i \in I_2\} \\
 &= \mu_1 g_1 + \mu_2 g_2
 \end{aligned}$$

Nous prendrons g comme origine; dans ces conditions, le moment d'inertie total $S(\mu_1)$ de N s'écrit :

$$\begin{aligned}
 S(\mu_1) &= \Sigma \{m(i)i^2 \mid i \in I\} \\
 &= \mu_1 \Sigma \{m_1(i)i^2 \mid i \in I_1\} + \mu_2 \Sigma \{m_2(i)i^2 \mid i \in I_2\} \\
 &= \mu_1 (S_1 + \|g_1\|^2) + \mu_2 (S_2 + \|g_2\|^2)
 \end{aligned} \tag{1}$$

Or g étant à l'origine, on a :

$$\mu_1 g_1 + \mu_2 g_2 = 0$$

d'où l'on déduit :

$$\begin{aligned}
 \mu_1 g_1 &= -\mu_2 g_2 = \mu_1 \mu_2 (g_1 - g_2) \\
 \mu_1 \|g_1\| &= \mu_2 \|g_2\| = \mu_1 \mu_2 \|g_2 - g_1\| = \mu_1 \mu_2 R
 \end{aligned} \tag{2}$$

De (1) et de (2), l'on déduit :

$$\begin{aligned}
 S(\mu_1) &= \mu_1 (S_1 + \mu_2^2 R^2) + \mu_2 (S_2 + \mu_1^2 R^2) \\
 &= \mu_1 S_1 + \mu_2 S_2 + \mu_1 \mu_2 R^2
 \end{aligned}$$

7.2.2. On a :

$$\left. \begin{aligned}
 C_1(\mu_1) &= \mu_1 (S_1 + \mu_2^2 R^2) \\
 C_2(\mu_1) &= \mu_2 (S_2 + \mu_1^2 R^2)
 \end{aligned} \right\} \tag{3}$$

7.2.3. L'équation $C_1(\mu_1)/C_2(\mu_1) = \rho$ s'écrit :

$$\mu_1 (S_1 + (1 - \mu_1)^2 R^2) - \rho (1 - \mu_1) (S_2 + \mu_1^2 R^2) = 0 \tag{4}$$

ou encore :

$$\mu_1^3 (1 + \rho) R^2 - \mu_1^2 (2 + \rho) R^2 + \mu_1 (S_1 + \rho S_2 + R^2) - \rho S_2 = 0 \tag{5}$$

C'est une équation du 3ème degré en μ_1 , de la forme $f(\mu_1) = 0$, qui a toujours au moins une racine comprise entre 0 et 1, puisque $f(0) = -\rho S_2 < 0$, $f(1) = S_1 > 0$.

L'équation $C_1(\mu_1)/C_2(\mu_1) = \rho$ (avec $0 \leq \mu_1 = 1 - \mu_2 \leq 1$) qui admet toujours une solution, peut en avoir trois si tel est le nombre de racines comprises entre 0 et 1, de l'équation $f(\mu_1) = 0$.

7.2.4. On a dans le cas de figure ci-jointe :

$$S_1 = S_2 = a^2 ; \|g_1 g_2\| = 2 ; R^2 = 4 ;$$

d'où l'on déduit d'après (1) et (2) que :

$$\begin{cases} S(\mu_1) = a^2 + 4 \mu_1 \mu_2 \\ C_1(\mu_1) = \mu_1 (a^2 + 4\mu_2^2) \\ C_2(\mu_1) = \mu_2 (a^2 + 4\mu_1^2) \end{cases} \quad (6)$$

L'équation (5) s'écrit alors, si $\rho = 1$:

$$f(\mu_1) = 8 \mu_1^3 - 12 \mu_1^2 + (2a^2 + 4) \mu_1 - a^2 = 0 \quad (7)$$

Cette équation qui admet, vu la symétrie entre I_1 et I_2 la solution $\mu_1 = 1/2$, s'écrit encore :

$$(2 \mu_1 - 1)(4 \mu_1^2 - 4 \mu_1 + a^2) = 0 \quad (8)$$

Le discriminant Δ' de $(4 \mu_1^2 - 4 \mu_1 + a^2)$ vaut $4(1 - a^2)$, donc :

- si $a^2 > 1$, l'équation $C_1(\mu_1)/C_2(\mu_1) = 1$ n'admet que la solution $\mu_1 = 1/2$
- si $a^2 \leq 1$, les deux racines de l'équation $(4 \mu_1^2 - 4 \mu_1 + a^2)$ appartiennent à l'intervalle $(0,1)$; l'équation $C_1(\mu_1)/C_2(\mu_1) = 1$ admet, outre la solution $\mu_1 = 1/2$, les solutions :

$$\mu_1' = (1/2)(1 - \sqrt{1 - a^2}); \mu_1'' = (1/2)(1 + \sqrt{1 - a^2}).$$

7.2.5. L'équation :

$$\frac{\mu_1(n)}{1-\mu_1(n)} = \left[\frac{C_1(\mu_1(n-1))}{\rho C_2(\mu_1(n-1))} \right]^\alpha \frac{\mu_1(n-1)}{1-\mu_1(n-1)},$$

s'écrit d'après (3) :

$$\frac{\mu_1(n)}{1-\mu_1(n)} = \left[\frac{\mu_1(n-1)}{1-\mu_1(n-1)} \right]^{1+\alpha} \left[\frac{S_1 + (1-\mu_1(n-1))^2 R^2}{\rho(S_2 + \mu_1^2(n-1) R^2)} \right]^\alpha; \quad (9)$$

ou encore, si $R^2 = 0$:

$$\frac{\mu_1(n)}{1-\mu_1(n)} = \left[\frac{\mu_1(n-1)}{1-\mu_1(n-1)} \right]^{1+\alpha} \left[\frac{S_1}{\rho S_2} \right]^\alpha \quad (10)$$

Or si $R^2 = 0$, l'équation $C_1(\mu_1)/C_2(\mu_1) = \rho$ s'écrit :

$$\frac{\mu_1 S_1}{(1-\mu_1) S_2} = \rho; \text{ ou : } \frac{\mu_1}{1-\mu_1} = \frac{\rho S_2}{S_1} = \left[\frac{S_1}{\rho S_2} \right]^{-1}. \quad (11)$$

On voit donc, en comparant (10) et (11), que si on prend $\alpha = -1$, quel que soit $\mu_1(0)$, $\mu_1(1)$ vérifie (11), i.e. est solution de $C_1(\mu_1)/C_2(\mu_1) = \rho$.

La formule (9) s'écrit alors :

$$\frac{\mu_1(n)}{1-\mu_1(n)} = \rho \frac{S_2 + \mu_1^2(n-1) R^2}{S_1 + (1-\mu_1(n-1))^2 R^2} \quad (12)$$

7.2.6. Dans le cas du nuage étudié au 4); (12) s'écrit, si $\rho = 1$:

$$\frac{\mu_1(n)}{1-\mu_1(n)} = \frac{a^2 + 4\mu_1^2(n-1)}{a^2 + 4(1-\mu_1(n-1))^2}$$

et l'on sait que l'équation :

$$\mu_1/(1-\mu_1) = (a^2 + 4\mu_1^2)/(a^2 + 4(1-\mu_1)^2)$$

admet la solution $\mu_1 = 1/2$, ainsi que, si $a^2 < 1$, les solutions :

$$\mu_1' = (1/2)(1 - \sqrt{1-a^2}) ; \quad \mu_1'' = (1/2)(1 + \sqrt{1-a^2})$$

Nous allons étudier la convergence de l'algorithme, au voisinage d'une des solutions de l'équation précédente.

Etude au voisinage de $\mu_1 = 1/2$:

Posons $\mu_1(0) = 1/2 - \varepsilon$, on a alors :

$$\left. \begin{aligned} \frac{\mu_1(1)}{1-\mu_1(1)} &= \frac{a^2 + 4(1/2 - \varepsilon)^2}{a^2 + 4(1/2 + \varepsilon)^2} = \frac{a^2 + 1 - 4\varepsilon + 4\varepsilon^2}{a^2 + 1 + 4\varepsilon + 4\varepsilon^2} \\ \mu_1(1) &= \frac{a^2 + 1 - 4\varepsilon + 4\varepsilon^2}{2(a^2 + 1) + 8\varepsilon^2} = 1/2 - \frac{2\varepsilon}{a^2 + 1 + 4\varepsilon^2} \end{aligned} \right\} \quad (13)$$

d'où :

$\mu_1(1)$ sera plus proche de $1/2$ que $\mu_1(0)$ si :

$$\frac{2|\varepsilon|}{a^2 + 1 + 4\varepsilon^2} < |\varepsilon| ;$$

i.e. si : $a^2 - 1 + 4\varepsilon^2 > 0$;

ou encore, puisque ε^2 est un infiniment petit du second ordre, si :

$$a^2 - 1 \geq 0 .$$

Si donc, $a^2 \geq 1$, auquel cas l'équation $C_1(\mu_1)/C_2(\mu_1) = 1$ n'a qu'une seule solution $\mu_1 = 1/2$ (*), l'algorithme précédent converge vers cette solution.

Par contre si $a^2 < 1$, l'algorithme précédent ne converge pas vers $1/2$, mais l'on sait qu'il existe alors deux autres solutions, μ_1' et μ_1'' , au voisinage desquelles on va maintenant étudier la stabilité de la convergence.

Etude au voisinage de $\mu_1 = (1/2)(1 + \varepsilon\sqrt{1-a^2})$, où $\varepsilon = 1$ ou -1 :

Posons $\mu_1(0) = (1/2)(1 + \varepsilon\sqrt{1-a^2} + u)$ où u est un infiniment petit. On a alors :

(*) Cette solution correspond à une racine de (7) de multiplicité 3, si $a^2 = 1$.

$$\begin{aligned}\frac{\mu_1(1)}{1-\mu_1(1)} &= \frac{a^2 + (1 + \epsilon \sqrt{1-a^2} + u)^2}{a^2 + (1 - \epsilon \sqrt{1-a^2} - u)^2} \\ &= \frac{2(1 + \epsilon \sqrt{1-a^2})(1+u) + u^2}{2(1 - \epsilon \sqrt{1-a^2})(1-u) + u^2} = \frac{A}{B}\end{aligned}$$

d'où l'on déduit :

$$\mu_1(1) = \frac{A}{A+B} = \frac{2(1 + \epsilon \sqrt{1-a^2})(1+u) + u^2}{4(1 + u \epsilon \sqrt{1-a^2}) + 2u^2}$$

En se bornant aux développements limités du premier ordre en u , il vient :

$$\begin{aligned}\mu_1(1) &\approx \mu_1 \times (1+u)/(1 + \epsilon u \sqrt{1-a^2}) \\ &\approx \mu_1 (1 + u(1 - \epsilon \sqrt{1-a^2})) \\ &\approx \mu_1 + (u(1 + \epsilon \sqrt{1-a^2})(1 - \epsilon \sqrt{1-a^2})/2) \\ &\approx \mu_1 + (u a^2/2),\end{aligned}$$

qui est donc plus proche de $\mu_1 = (1 + \epsilon \sqrt{1-a^2})/2$, que ne l'est :

$$\mu_1(0) = (1 + \epsilon \sqrt{1-a^2} + u)/2 = \mu_1 + (u/2),$$

car a^2 est inférieur à 1.

$$\begin{array}{ccccccc} \mu_1 & & \dots & & \mu_1(2) & & \mu_1(1) & & \mu_1(0) \\ \times & & & & \times & & \times & & \times \end{array}$$

L'algorithme converge donc vers $(1/2)(1 + \epsilon \sqrt{1-a^2})$, et l'on peut remarquer que $\mu_1(n)$ est toujours du même côté que $\mu_1(0)$ par rapport à $\mu_1 = (1/2)(1 + \epsilon \sqrt{1-a^2})$, position limite atteinte quand $n \rightarrow \infty$; de plus la distance à la position limite décroît comme une suite géométrique de raison a^2 (cf. figure).

7.2.7. L'équation (12) qui s'écrit :

$$\mu_1(n) = \rho \frac{[S_2 + \mu_1^2(n-1)R^2]}{S_1 + \rho S_2 + R^2(\rho \mu_1^2(n-1) + (1 - \mu_1(n-1))^2)} \quad (14)$$

est de la forme :

$$\mu_1(n) = g(\mu_1(n-1)) \quad (15)$$

Recherchant une solution de $C_1(\mu_1)/C_2(\mu_1) = \rho$, équation qui est équivalente à $\mu_1 = g(\mu_1)$, d'après (3), (14) et (15), on résout donc cette dernière équation par itérations successives.

Si $\hat{\mu}_1$ désigne une solution de $\mu_1 = g(\mu_1)$, appartenant à l'intervalle $(0,1)$ (on sait d'après 3°, qu'il existe une telle solution), et si l'on est au voisinage de $\hat{\mu}_1$, l'algorithme précédent convergera vers $\hat{\mu}_1$ si $|g'(\hat{\mu}_1)| < 1$ et ne convergera pas vers $\hat{\mu}_1$ si $|g'(\hat{\mu}_1)| > 1$ (i.e. $\hat{\mu}_1$ est une solution stable si $|g'(\hat{\mu}_1)| < 1$, instable si $|g'(\hat{\mu}_1)| > 1$).
Posons :

$$N(\mu_1) = \rho(S_2 + \mu_1^2 R^2) ;$$

$$D(\mu_1) = S_1 + \rho S_2 + R^2(\rho \mu_1^2 + (1 - \mu_1)^2).$$

Avec ces notations, on a :

$$g(\mu_1) = N(\mu_1)/D(\mu_1)$$

$$g(\hat{\mu}_1) = N(\hat{\mu}_1)/D(\hat{\mu}_1) = \hat{\mu}_1$$

$$g'(\mu_1) = \frac{N'}{D} - \frac{N}{D} \frac{D'}{D}$$

$$\begin{aligned} g'(\hat{\mu}_1) &= (N'(\hat{\mu}_1) - \hat{\mu}_1 D'(\hat{\mu}_1))/D(\hat{\mu}_1) \\ &= \hat{\mu}_1 (N'(\hat{\mu}_1) - \hat{\mu}_1 D'(\hat{\mu}_1))/N(\hat{\mu}_1); \end{aligned}$$

d'où il vient, en tenant compte des valeurs de N' et D' :

$$N'(\hat{\mu}_1) = 2 \rho \hat{\mu}_1 R^2; \quad D'(\hat{\mu}_1) = 2 R^2(\rho \hat{\mu}_1 - (1 - \hat{\mu}_1)),$$

$$\begin{aligned} N'(\hat{\mu}_1) - \hat{\mu}_1 D'(\hat{\mu}_1) &= 2 \hat{\mu}_1 R^2(\rho(1 - \hat{\mu}_1) + (1 + \hat{\mu}_1)) \\ &= 2 \hat{\mu}_1 (1 - \hat{\mu}_1) R^2(1 + \rho). \end{aligned}$$

La condition de stabilité (i.e. de convergence) $|g'(\hat{\mu}_1)| < 1$ s'écrit donc, puisque $g'(\hat{\mu}_1) > 0$:

$$g'(\hat{\mu}_1) = \frac{2 \hat{\mu}_1^2 (1 - \hat{\mu}_1) R^2 (1 + \rho)}{\rho(S_2 + \hat{\mu}_1^2 R^2)} < 1$$

Dans le cas du nuage étudié en 4° et 6°, si $a^2 > 1$, on a $\hat{\mu}_1 = 1/2$ comme seule solution, et puisque $R^2 = 4$, $S_2 = a^2$, $\rho = 1$, on a :

$$g'(\hat{\mu}_1) = 2/(1 + a^2),$$

qui est bien inférieur à 1 si $a^2 > 1$.

Par contre pour $a^2 < 1$, $g'(\hat{\mu}_1) > 1$, il n'y a pas convergence vers 1/2; tandis que, pour la racine $\hat{\mu}_1 = (1/2)(1 + \varepsilon \sqrt{1 - a^2})$, on a :

$$g'(\hat{\mu}_1) = \frac{2(1 + \varepsilon \sqrt{1 - a^2}) a^2}{a^2 + (1 + \varepsilon \sqrt{1 - a^2})^2} = a^2$$

qui est bien inférieur à 1.

Les solutions $(1/2)(1 + \sqrt{1 - a^2})$ et $(1/2)(1 - \sqrt{1 - a^2})$ sont stables. On retrouve bien les résultats du § 7.2.6.

Remarques :

1) $g'(\hat{\mu}_1)$ étant toujours positif, s'il y a convergence vers $\hat{\mu}_1$, et si $\mu_1(0)$ est assez proche de $\hat{\mu}_1$, la suite $\mu_1(n)$ est toujours située d'un même côté, par rapport à $\hat{\mu}_1$.

2) Au lieu de calculer $g'(\hat{\mu}_1)$, on aurait pu, comme au § 7.2.6, faire un développement limité au voisinage de $\hat{\mu}_1$.

7.2.8. Comme nous l'avons annoncé au début du paragraphe 7, il faut prendre garde qu'en analyse des correspondances la métrique dont

est muni l'espace $E = R_J$ est fonction du centre de gravité du nuage I. Ainsi dans les formules des paragraphes 7.2.1 et 7.2.2 les quantités S_1 , S_2 , R^2 sont elles-mêmes des fonctions de μ_1 : de façon précise les p_j dépendent linéairement de μ_1 , donc les $(1/p_j)$ coefficients de la métrique du χ^2 sont des fonctions homographiques de μ_1 (inverses de fonctions linéaires) et S_1 , S_2 , R^2 sont des fonctions rationnelles combinaisons de ces fonctions homographiques. Il en résulte que l'équation du paragraphe 7.2.3 devient une équation algébrique en μ_1 de degré généralement supérieur à 3. Toutefois dans la 4^e question l'espace E, étant une droite, est toujours muni de la même métrique, à un coefficient de proportionnalité près qui ne change rien à nos calculs, lesquels s'appliquent donc sans changement. Ce cas où E est une droite pourra paraître peu réaliste : poser $E = R_J \approx R$, implique en effet Card J = 1. Mais il faut noter d'abord que le nuage des p_j^i appartient à l'hyperplan de R_J , lieu des mesures de masse totale 1 : ce qui permet de poser Card J = 2. De plus nos calculs valent encore, quel que soit Card J, pourvu que les p_j^i soient tous sur une même droite (sous-espace affiné de dim. 1) de R_J ; c'est-à-dire si le tableau des données est de rang 2. Cette dernière situation est, elle, d'un réalisme acceptable pour un contre-exemple tel que celui étudié en 6°. Pour une autre raison le choix de α proposé en 5° vaut aussi en analyse des correspondances ; c'est qu'il s'agit justement du cas où, g étant indépendant de μ_1 , la métrique du χ^2 est toujours la même (cf supra § 6). On peut donc reprendre sans changement le 6°. Il n'en n'est évidemment pas de même du 7° : la situation étant comme nous l'avons dit bien plus complexe du fait de la mobilité du centre g. Mais il nous suffira d'avoir étudié en détail en 4° et 6° un contre-exemple d'instabilité valable dans le cas de l'analyse des correspondances.

7.2.9. Remarque sur la convergence de l'algorithme étudié aux questions 5° et 7° :

Les équations (12) et (14) s'écrivent en posant $\mu_1(n) = y$,

$\mu_1(n-1) = x$:

$$\frac{y}{1-y} = \rho \frac{S_2 + R^2 x^2}{S_1 + (1-x)^2 R^2} \quad (12')$$

$$y = \rho \frac{S_2 + R^2 x^2}{S_1 + \rho S_2 + (\rho x^2 + (1-x)^2) R^2} = g(x), \quad (14')$$

$y = g(x)$ étant toujours compris dans l'intervalle (0,1), les solutions de l'équation $x = g(x)$ (qui se réduit à l'équation du 3^e degré $f(x) = 0$, donnée par (5), où l'on pose $\mu_1 = x$) appartiennent à cet intervalle.

Par ailleurs, on a, en désignant par $D(x)$ le dénominateur de (14') :

$$y' = g'(x) = \frac{2 \rho R^2}{D^2(x)} (-R^2 x^2 + (S_1 - S_2 + R^2) x + S_2);$$

d'où l'on déduit que y' s'annule pour 2 valeurs x' et x'' telles que $x' \leq 0 < 1 \leq x''$, y' étant positif dans l'intervalle $(0,1)$; $g(x)$ est donc croissante dans cet intervalle. De plus, on a $g(0) > 0$, $g(1) < 1$, donc :

$$g(x) > x, \text{ si } x = 0;$$

$$g(x) < x, \text{ si } x = 1.$$

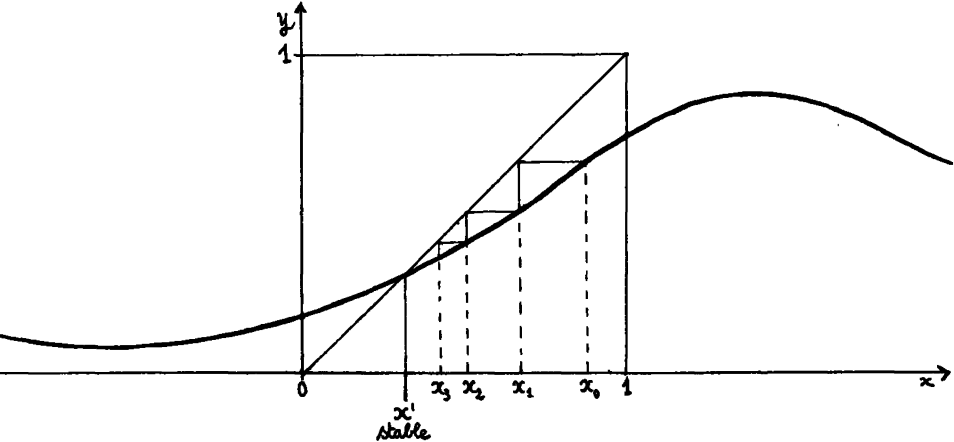
Des résultats précédents, l'on déduit :

a) que si l'équation $x = g(x)$ admet une seule racine, x' , $g'(x') < 1$, et l'algorithme proposé au 5° converge vers x' .

b) que si l'équation $x = g(x)$ admet 3 racines x' , x'' , x''' ($x' < x'' < x'''$), les valeurs x' et x''' sont stables ($g'(x') < 1$; $g'(x''') < 1$), tandis que x'' est instable ($g'(x'') > 1$).

Si donc l'on prend comme point de départ de l'algorithme proposé au 5° une valeur x_0 inférieure à x'' , il y a convergence vers x' ; si au contraire x_0 est supérieur à x'' , il y a convergence vers x''' .

Les résultats précédents sont évidents graphiquement, comme le montrent les figures ci-après.



Etude graphique de la convergence de l'algorithme

- ci-dessus, (a) : $x = g(x)$ a une solution x' ;
 ci-dessous, (b) : $x = g(x)$ a trois solutions, x' , x'' , x''' .

