Méthode des réseaux en analyse de données, application à l'analyse de concordance

Jean-Marie Tricot

Département d'Econométrie

ÉCONOMIQUES ET SOCIALES

Professeur Jean-Pierre SCHELLHORN

Méthode des réseaux en analyse de données, application à l'analyse de concordance

THÈSE

présentée à la Faculté des sciences
économiques et sociales
de l'Université de Genève
pour obtenir le grade de docteur ès sciences
économiques et sociales,
mention économétrie et statistique

par

Jean-Marie TRICOT

de Genève

Thèse Nº 359

GENÈVE
Imprimerie Nationale
1990

Méthode des réseaux en analyse de données, application à l'analyse de concordance

Thèse présentée à la Faculté des sciences économiques et sociales de l'Université de Genève

par
Jean-Marie Tricot

pour l'obtention du grade de
Docteur ès sciences économiques et sociales
mention économétrie et statistique

Membres du jury de thèse

M. Jean-Pierre Schellhorn
professeur, Genève, directeur de thèse

M. Pietro Balestra
professeur, Genève, président du jury

M. Franz Streit
professeur, Genève

M. Elvezio Ronchetti
professeur, Genève

M. Jacques Royer
professeur, Genève

M. Bernard Van Cutsem
professeur, Grenoble

These No. 359

La Faculté des sciences économiques et sociales, sur préavis du jury, a autorisé l'impression de la présente thèse, sans entendre, par là, émettre aucune opinion sur les propositions qui s'y trouvent énoncées et qui n'engagent que la responsabilité de leur auteur.

Genève, le 29 juin 1990

Le doyen:
Paolo Urio

TABLE DES MATIERES

INTRODUCTION

Dans les différents domaines de la statistique descriptive, les données se présentent sous forme de nuages de points ; sur ceux-ci, on est souvent amené à faire des études de proximité ou, plus généralement, de similarité, permettant d'effectuer des analyses de structure.

Il en est ainsi en analyse de concordance où il s'agit d'apprécier le degré d'accord entre d observateurs évaluant le même ensemble de n sujets au moyen d'une échelle de valeurs possibles prises par une variable (on peut généraliser le problème à plusieurs variables).

Ce degré d'accord se traduit en terme de proximité ou de similarité des observateurs et dans un premier temps, on rappelle que l'étude de l'accord peut se ramener, dans certains cas, à une étude de corrélation des composantes d'un vecteur multidimensionnel nxd d'évaluations $(X_1 , \ldots, X_d)$. Mais, un cas important est celui d'une variable qualitative nominale. Dans un tel cas, la notion d'accord, sur chaque sujet, peut être associée à un certain degré de "cohésion" d'un nuage de points pondérés.

Ainsi, on voit qu'en analyse de concordance, l'analyse de données, telle qu'elle se pratique par l'analyse de la structure d'un nuage de points, joue un très grand rôle.

Cette notion de "cohésion" d'un nuage de points, je l'ai abordée, dans un premier temps, comme on fait en analyse en composantes principales, à l'aide d'une variance. C'est une approche nouvelle en analyse de concordance.

La notion de dispersion, ici, dispersion des évaluations, est plutôt liée à celle de désaccord ; symétriquement, j'ai voulu construire une méthode fondée cette fois sur la notion d'accord, ce qui fait intervenir la notion de concentration dans un modèle de nuage de points.

Voilà pourquoi, dans une première partie, il apparaît très important de préciser et de résoudre le problème d'une analyse de la structure d'un nuage de points à l'aide d'un indice de concentration.

J'ai utilisé celui de C., Tricot (1971) qui a l'avantage de ne pas faire appel à la notion de distance.

Cette étude de structure d'un nuage fondée sur la notion de concentration, permet de résoudre, dans une deuxième partie, un problème d'analyse de concordance.

Avant de poursuivre pour décrire les thèmes de cet ouvrage, je désire remercier ici le professeur Claude Tricot de m'avoir ouvert la voie dans cette approche théorique si fructueuse en elle-même et guidé tout au long de ce travail.

Exposons maintenant, plus précisément, ce qui fait l'objet de la première partie.

L'un des problèmes de l'analyse de données est celui de la recherche des structures de ces données par la réduction de la grande quantité d'informations qu'elles contiennent.

C'est pourquoi on est amené à projeter les données dans un espace de dimension inférieure à la dimension d'origine.

Comme le souligne Huber (1987), se pose alors les questions du rôle et du choix d'une projection et par conséquent, d'un indice de projection : ces questions interviennent dans le cadre d'une analyse qui a été appelée "Projection Pursuit" par Kruskal à la fin des années 60.

Lorsqu'on parle d'indice de projection, on s'intéresse à l'aspect analytique d'une "Projection Pursuit" plutôt qu'aux méthodes purement graphiques de recherche de structures, basées sur des techniques informatiques interactives.

J'ai tout d'abord montré qu'un certain nombre de méthodes d'analyse de données classiques pouvaient être présentées sous l'angle de "Projection Pursuit".

Ainsi, la régression linéaire, l'analyse factorielle discriminante, l'analyse en composantes principales et l'analyse d'un nuage par l'entropie, sont autant de méthodes destinées

à construire un sous-espace de $\mathbb{R}^p$ sur lequel on projette le nuage Ω des n points à p composantes considéré, ce qui fait apparaître une structure particulière de ce nuage : une telle structure pourra être considérée comme proche d'une structure définie a priori ou encore de la structure initiale dans $\mathbb{R}^p$.

Les indices de projection utilisés dans ces méthodes, sont essentiellement des distances interpoints ou, dans le cas de l'entropie, une mesure de normalité en considérant que le tableau de données pxn est un échantillon d'une variable aléatoire multidimensionnelle.

Ces différents exemples de "Projection Pursuit" illustrent bien la manière dont s'opère la projection : ils mettent l'accent sur les problèmes d'observation du nuage dans $\mathbb{R}^p$ et sur les propriétés de l'indice de projection.

Puis, sont énoncées certaines propriétés d'un indice de projection. Elles peuvent garantir l'invariance des sous-espaces d'observation des structures recherchées pour les transformations effectuées sur Ω. Elles peuvent également apporter des simplifications dans les techniques de calcul.

J'ai présenté ensuite un nouvel indice ayant ce grand avantage de ne faire appel, comme je l'ai signalé plus haut, ni à la notion de distance interpoints ni à celle de non-normalité.

Ne pas faire appel à la notion de distance permet notamment de réduire la trop grande contribution des points aberrants dans la recherche du sous-espace. D'un autre point de vue, on se soustrait au problème de concevoir précisément un degré de non-normalité puisqu'on laisse de côté toute recherche de structure normale.

Ce nouvel indice (appelé aussi coefficient de concentration) apparaît comme un "indicateur" d'une répartition des points de Ω dans l'espace : dans une direction de projection donnée, l'échantillon des points projetés pourra être considéré comme généré par une distribution de Dirac si l'indice est maximum et par une distribution uniforme si l'indice est minimum.

On recherche donc au moyen de la "Projection Pursuit" utilisant un tel indice, des sous-espaces de $\mathbb{R}^p$ ou la projection du nuage soit de concentration minimum ou encore de concentration maximum.

A ce niveau, les techniques sont assez délicates. Pour déboucher sur des problèmes d'optimisation classiques, il est nécessaire d'exprimer l'indice sous une forme analytique approchée et l'on obtient ainsi une fonction objectif.

A titre d'exemple, j'ai étudié, à la fin de ce chapitre, certains problèmes de projection en dimension 2, ce qui amène comme en régression linéaire classique, à décrire le nuage Ω

en dimension 1.

Dans une deuxième partie, j'ai appliqué ces idées sur l'analyse de la structure d'un nuage, aux problèmes spécifiques de l'analyse de concordance.

J'ai donc été conduit à comparer les notions d'accord, de dispersion et de concentration.

Tout d'abord, j'ai situé le cadre d'une analyse de concordance en donnant plusieurs exemples d'une telle analyse, tirés de la littérature, suivant que les évaluations sont des valeurs prises par une variable quantitative ou une variable qualitative ordinale ou encore une variable qualitative nominale.

Dans ce dernier cas, l'indice d'accord généralement utilisé est la mesure Kappa pondérée dont j'ai rappelé la définition en indiquant, à ce propos, comment est formulée la notion de désaccord et d'après cela, comment est testé ce désaccord au vu de la mesure obtenue.

J'ai repris ce cas d'une variable qualitative nominale et j'ai introduit la notion de fonction d'indice qui associe, à toute partition du nombre d d'observateurs une mesure du degré d'accord sur chaque sujet.

Un ordre sur l'ensemble des partitions de d est déterminé à l'aide de l'ordre des valeurs prises par une telle fonction d'indice.

J'ai montré que le coefficient de Cartwright, dont dépend la mesure Kappa, est la valeur prise par une fonction d'indice et que cette valeur induit un ordre sur les partitions de d et de là, définit une forme de progression dans l'accord qui n'est pas nécessairement la meilleure car elle est fondée sur les accords deux à deux des observateurs à propos de chaque sujet.

Ceci donne lieu à la définition d'un indice d'accord construit de la manière suivante :
L'accord est une relation d'équivalence qui induit une partition des observateurs à propos de chaque sujet.

Lorsque l'accord n'est pas parfait (plus d'un élément dans la partition), la variance des effectifs de ces éléments induit un ordre sur les partitions de d qui répond à de bons critères d'accord.

De cette façon de procéder, on déduit immédiatement, par normalisation (catégorisation des β_k), un nouvel indice d'accord sur [0, 1], basé sur une notion de dispersion des observateurs en groupes d'observateurs.

J'ai donné ensuite, grâce à une représentation des observa-

teurs par un nuage de points sur la droite, un autre indice d'accord basé, cette fois, sur la notion de concentration.

L'introduction d'un coefficient correcteur sur l'un et l'autre des deux indices dont on vient de parler, permet, lorsqu'il sont tous les deux normalisés, de les comparer entre eux.

Sortant de la statistique descriptive, les sujets représentant maintenant un échantillon, j'ai développé un ensemble de tests d'hypothèses paramétriques permettant de décider si l'on peut admettre l'existence d'un accord entre les observateurs.

Et pour finir, j'ai réalisé une description spatiale, à l'aide, de nouveau, d'un nuage de points dans le plan, description d'une confrontation du comportement individuel d'un observateur, au comportement général des autres. A chaque observateur on peut ainsi attribuer une ligne polygonale représentative de son nuage de points et l'on peut comparer les d lignes polygonales obtenues, afin de détecter si, parmi les observateurs, certains ont une conduite aberrante.

Une théorie statistique ne saurait se passer des instruments informatiques lui permettant d'être verifiée dans des applications. Ainsi, on trouvera deux programmes aux annexes A 2 et A 3 .

Enfin, j'ai eu la chance de me voir confier une étude importante par l'OMS, portant sur des observations recueillies dans 42 pays ce qui m'a permis d'utiliser certaines des méthodes qui sont décrites précédemment et de présenter quelques résultats empiriques (annexe A 4). Un diagnostic est, en effet, une évaluation portée par un observateur sur un sujet.

Au terme de cette introduction, je voudrais remercier le professeur J.-P. Schellhorn d'avoir accepté de diriger cette recherche et de m'avoir relancé au cours du travail grâce à de précieuses références ou des conseils judicieux, notamment pour la synthèse générale.

Je remercie également le professeur P. Balestra d'avoir accepté de présider le jury : son expérience extrêmement variée lui permet une excellente appréciation de problèmes fort divers et c'est dans l'atmosphère très "conviviale" de "son" département que s'est inscrit mon travail.

Je remercie le professeur E. Ronchetti qui m'a si utilement indiqué la voie spécifique dans laquelle je me suis engagé en analyse de données ; au fond, il a "situé" la première partie de mon travail.

Je remercie le professeur J. Royer qui m'a toujours appris, durant mes études et jusqu'à cette thèse, à développer l'aspect pratique de la science sous-tendu par de fortes théories : il faut appliquer et pas seulement rêver.

Je remercie le professeur F. Streit de s'être toujours intéressé, à chacune de nos rencontres, aux problèmes qu'il m'a fallu traiter et à qui je dois bon nombre de lectures adéquates.

Je remercie enfin le professeur B. Van Cutsem de l'Université de Grenoble I d'avoir accepté de faire partie de mon jury et d'avoir bien voulu apporter ainsi la caution scientifique si précieuse d'un statisticien renommé.

Je voudrais terminer en exprimant ma gratitude à la section de mathématiques et au département d'économétrie de l'Université de Genève. J'ai pu trouver dans ces deux cadres une ambiance intellectuelle enrichissante, beaucoup de possibilités de contacts avec l'étranger et plusieurs bibliothèques de très haut niveau.

Mes remerciements sont également adressés à Madame Angela Bianchet et à Mademoiselle Dominique Eyer qui ont assuré avec talent ce long travail de dactylographie très fastidieux, soyons réalistes.

PREMIERE PARTIE

METHODE DES RESEAUX EN ANALYSE DE DONNEES

CHAPITRE 1

Projections et Exemples

1.1 PROBLEME MULTIDIMENSIONNEL ET NOTATIONS

L'analyse statistique de la répartition de n points dans l'espace $\mathbb{R}^p$ se rapporte à l'étude d'un tableau de données p x n :

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{p1} & \cdots & x_{pn} \end{pmatrix}$$

La jème colonne du tableau est la représentation d'un point d'un nuage Ω. Schématiquement, comme on lit dans Cailliez et Pages (1976), cette jème colonne est un vecteur de réponses à un questionnaire comprenant p questions.

Il y a n points notés $\underline{x}_j$:

$$\underline{x}_j = (x_{1j}, \cdots, x_{pj})'$$

La ième ligne du tableau est la représentation d'un échantillon de valeurs prises par une variable réelle soit quantitative soit qualitative : une telle variable est une application de Ω dans un espace d'observations O. Si l'on peut munir O d'une structure de σ-algèbre, une telle variable pourra être appelée "variable aléatoire".

Il y a p variables et les échantillons sont notés vectoriellement $\underline{x}^i$:

$$\underline{x}^i = (x_{i1}, \cdots, x_{in})'$$

Selon les termes de Jones et Sibson (1987), on s'intéresse ici à une analyse exploratoire de données.

On attribue à ces données une structure pour l'analyse : on les considère comme un ensemble de n vecteurs de dimension p.

Toujours selon les termes de Jones et Sibson (1987), on espère pouvoir trouver et décrire, en terme de tendance, deux qualités du nuage Ω :

a) L'hétérogénéité : c'est l'aspect classification ou discrimination.
b) L'homogénéité : c'est l'aspect analyse factorielle qui s'intéresse aux zones d'occupation privilégiées.

A cette fin, on se réfère à la méthode dite "Projection

homogènes d'une transformation du nuage par projection sur un sous-espace de $\mathbb{R}^p$; ceci se justifie par le fait qu'on ne peut pas facilement donner une description exhaustive d'un nuage plongé dans un espace de dimension p trop élevée.

Ainsi, deux problèmes se posent :

1) Une recherche d'une projection du nuage.
2) Une recherche des structures du nuage.

Ces deux problèmes sont liés : pour trouver de "bonnes" structures, il faut trouver une bonne projection qui les fasse apparaître.

Il y a deux démarches principales possibles :

A : Avoir une idée a priori de ce que sont de bonnes structures et trouver en fonction de cela une projection.

B : Avoir une idée a priori de ce qu'est une bonne projection et faire l'analyse des structures induites.

1.2 LA METHODE "PROJECTION PURSUIT"

La recherche d'une projection liée à certaines structures du nuage a été effectuée de deux manières dans la littérature :

1) Par une procédure interactive à l'aide d'un système informatique graphique (voir par exemple Launer et

Une "inspection visuelle" d'un ensemble d'hyperplans de projection inclus dans $\mathbb{R}^p$, permet de trouver des répartitions de Ω en classes et en hypersurfaces.

2) Par l'optimisation d'un indice de projection autrement appelé "critère" en analyse de dissimilarités. D'après Critchley (1988), les qualités spéciales d'un tel indice sont directement liées au type de représentation du nuage obtenu, c'est-à-dire aux structures obtenues.

Par la suite, on va s'intéresser à cette deuxième procédure de recherche qu'on appellera "exploration par indice de projection".

1.3 QUELQUES EXEMPLES D'EXPLORATIONS PAR INDICE DE PROJECTION

1.3.1 L'analyse de régression linéaire classique

Dans $\mathbb{R}^p$ on sélectionne une variable, par exemple la première, et afin de la distinguer des autres, on note y_j sa jème valeur prise dans l'échantillon $\{x_{11}, \ldots, x_{1n}\}$:

$$y_j = x_{1j}$$

Ainsi, dans les nouvelles notations,

$$X = \begin{pmatrix} y_1 & \cdots & y_n \\ x_{21} & \cdots & x_{2n} \\ \vdots & & \vdots \\ x_{p1} & \cdots & x_{pn} \end{pmatrix}$$

On prend dans cet exemple la démarche A :

On désire projeter Ω sur un sous-espace de dimension p-1 faisant apparaître la structure suivante : une liaison affine entre la première variable et les p-1 autres variables.
On demande de plus que le nouveau nuage projeté ait le même centre de gravité que Ω.

Le jème point du nouveau nuage doit donc appartenir au sous-espace affine de $\mathbb{R}^p$ défini par un vecteur et un point :

- un vecteur, orthogonal à ce sous-espace affine, dont on peut évidemment fixer une composante :

$$\beta = \begin{pmatrix} -1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \qquad (\beta_i \in \mathbb{R})$$

- le point :

$$g = \frac{1}{n} XS \qquad (S = (1, \ldots, 1)')$$

$$z_j = \begin{pmatrix} y_j \\ x_{2j} \\ \vdots \\ x_{pj} \end{pmatrix}$$

sur l'espace de dimension p-1 qui nous intéresse est donc nécessairement le point :

$$\hat{z}_j = \begin{pmatrix} \bar{y} + \sum_i \beta_i (x_{ij} - \bar{x}^i) \\ x_{2j} \\ \vdots \\ x_{pj} \end{pmatrix}$$

où $\bar{y} = \frac{1}{n} \sum_j y_j$ et $\bar{x}^i = \frac{1}{n} \sum_j x_{ij}$

$\hat{z}_j$ appartient bien à l'espace de projection.

L'indice de projection $I(\beta)$ est défini par "une distance interpoint" :

$$I(\beta) = \sum_j \| z_j - \hat{z}_j \|^2$$

où $\| \ \|^2$ désigne le carré de la norme euclidienne.

Cependant, on remarque qu'il n'est pas nécessaire de parler de norme, mais seulement de forme quadratique de matrice identité.

rapport aux β_i.

Remarque 1
==========

La projection utilisée n'est pas une projection orthogonale sauf si les points sont déjà dans l'hyperplan de projection. ■

Remarque 2
==========

On a obtenu dans cet exemple une structure particulière sur Ω par une méthode de projection. Cependant, si $p \geq 4$, il devient difficile, voir impossible, d'apprécier visuellement la projection : la projection ne devient plus "visible". ■

1.3.2 L'analyse en composante principale

On prend dans cet exemple la démarche B.

On désire projeter le point $\underline{x}_j$ de Ω sur le point $\hat{\underline{x}}_j$ appartenant à un hyperplan W de $\mathbb{R}^p$ de telle sorte que l'inertie de Ω autour de W soit minimisée.

Une telle projection est considérée comme une bonne projection par référence à une proximité au sens physique de Ω à sa projection.

étape ultérieure à cette projection.

L'indice de projection est défini par une distance interpoint :

$$\tilde{I} = \sum_{j=1}^{n} \| \underline{x}_j - \hat{\underline{x}}_j \|^2$$

où $\| \ \|^2$ est une forme quadratique de matrice M définie positive (on suppose, pour simplifier l'exposé de la méthode, que les points ne sont pas pondérés).

Caractéristiques de la projection (justifiant le terme "inertie")

1) La projection est orthogonale :

Elle projette orthogonalement des points transformés des points de Ω par l'application $X \longmapsto M^{1/2}X$.

Exemple : si $M^{1/2}$ est diagonale, c'est une matrice de changement d'échelle sur les axes initiaux qui peut se traduire par une standardisation des variables. ■

2) W est un hyperplan affine passant par le centre de gravité de Ω. ■

Du fait de l'orthogonalité de la projection, le théorème de Pythagore nous donne :

$$\tilde{I} = \sum_j \| \underline{x}_j \|^2 - \sum_{k=1}^{q} \lambda_k$$

où les λ_k sont q valeurs propres associées aux q vecteurs propres de $M^{1/2}X\ (I - \frac{SS'}{n})\ X'M^{1/2}$ constituant une base de W.

Pour q fixé, $\tilde{I}$ est minimisé en prenant les q plus grandes valeurs propres de $M^{1/2}X\ (I - \frac{SS'}{n})\ X'M^{1/2}$ celles dont les vecteurs propres associés donnent la direction de W.

W passe par le point $\frac{1}{n}$ XS comme dans l'exemple a).

En ce qui concerne les structures de Ω observées sur une projection en dimension 2, on note que :

- la réunion des points en classes ne découle que de l'idée de proximité au sens $\| \ \|^2$.

- les proximités entre points et axes factoriels, ou éventuellement entre points et composantes principales, ont une interprétation qui présuppose un modèle multidimensionnel gaussien sur les variables.

Remarque
========

Minimiser $\tilde{I}$ revient, de façon équivalente, à maximiser la somme des variances des composantes principales puisque ces variances sont précisément les valeurs propres λ_k ci-dessus.

L'ensemble des coordonnées de ces composantes principales forment la projection de Ω sur chacun des axes factoriels de la base de W.

Ainsi la minimisation de $\tilde{I}$ consiste à rechercher la plus grande dispersion des points du nuage sur chacun des axes d'un système d'axes orthogonaux.

Une telle dispersion s'interprète comme une somme de distance à un centre de gravité : il s'agit encore d'une distance interpoint. ■

1.3.3 L'analyse factorielle discriminante

Cet exemple, bien que proche du précédent, a l'avantage d'associer les deux démarches A et B dans la même analyse.

On sait que l'aspect classification est éliminé du problème puisque l'on suppose connaître déjà une séparation de Ω en classes d'individus.

On s'intéresse à trouver des règles d'affectation relativement simples, à l'aide de fonctions discriminantes, permettant de définir l'appartenance à une classe d'un individu de Ω ou de tout autre individu venant s'intégrer à Ω (cadre prévisionnel).

Comme on lit dans Diday (1982), on désire "éviter des calculs

d'information concernant le problème posé".

Ainsi une méthode prospective s'impose. Elle consiste à rechercher la projection qui rende le mieux possible compte de la structure de Ω en classes : c'est une démarche de type A; ou, de façon équivalente, elle consiste à rechercher la meilleure projection de chacune des classes : c'est alors une démarche de type B.

Lorsque l'on procède suivant B, on définit les classes par leur centre de gravité et l'on fait une analyse en composante principale sur le tableau $\tilde{X}$ (p x $\tilde{n}$) de ces centres de gravité préalablement "sphérés" : c'est-à-dire que l'on prend $M = \tilde{n}\,[\tilde{X}\,(I - \frac{SS'}{\tilde{n}})\,\tilde{X}']^{-1}$. On dit parfois que l'on met sur $\mathbb{R}^p$ la distance de Mahalanobis.

Lorsque l'on procède suivant A, on cherche à maximiser la variance interclasse des points du nuage projetés sur chaque axe d'un système d'axes orthogonaux; ou, ce qui revient au même, à minimiser la variance intraclasse.

Les indices de projections sont encore, comme dans les exemples précédents, des distances interpoints.

1.3.4 L'analyse du nuage par l'entropie

On prend ici une démarche du type A modifiée : au lieu de

structures "inintéressantes" ce qui est une manière par élimination de trouver les bonnes.

Une structure multidimensionnelle normale des p variables est décelable rapidement à l'aide d'un calcul de l'entropie sur une projection unidimensionnelle des points de Ω (voir Huber, 1985). C'est ce que l'on va préciser maintenant.

Soit α une direction de projection (α est un vecteur p x 1).

Prenons $\|\alpha\| = 1$.

Le tableau X est assimilable à un n-échantillon extrait d'une variable aléatoire multidimensionnelle et, dans ce cas, en projection, le vecteur $(z_1, \ldots, z_n)' = \alpha'X$ est assimilable à un n-échantillon $(z_1, \ldots, z_n)_\alpha$ extrait d'une variable aléatoire Z_α. On suppose que sa distribution est absolument continue sur $(-\infty, +\infty)$, de densité f_α, et qu'elle possède un moment d'ordre 2.

De l'entropie de Z_α : $-\int_{-\infty}^{+\infty} f_\alpha(z) \log \frac{1}{f_\alpha(z)}\, dz$

on peut déduire l'indice suivant défini pour toute projection dans la direction α :

$$I(\alpha) = -\int_{-\infty} f_\alpha(z) \log \frac{1}{f_\alpha(z)} dz + \log \sqrt{2\pi e \text{ Var } Z_\alpha}$$

qui est une simple standardisation de l'entropie par affinité.

Proposition

Soit $\mu = E(Z_\alpha)$

1) $I(\alpha) = 0 \iff f_\alpha(z) = \frac{1}{\sqrt{2\pi \text{ Var } Z_\alpha}} e^{-\frac{(z-\mu)^2}{2 \text{ Var } Z_\alpha}}$

2) $I(\alpha) > 0$ si $f_\alpha(z) \neq \frac{1}{\sqrt{2\pi \text{ Var } Z_\alpha}} e^{-\frac{(z-\mu)^2}{2 \text{ Var } Z_\alpha}}$

En effet :

on pose : $\varphi_\alpha(z) = \frac{1}{\sqrt{2\pi \text{ Var } Z_\alpha}} e^{-\frac{(z-\mu)^2}{2 \text{ Var } Z_\alpha}}$

on a :
$$\int_{-\infty}^{+\infty} f_\alpha(z) \log \varphi_\alpha(z) dz = -\log \sqrt{2\pi \text{ Var } Z_\alpha} - \frac{1}{2}$$
$$= -\log \sqrt{2\pi e \text{ Var } Z_\alpha}$$

Donc :
$$I(\alpha) = -\int_{-\infty}^{+\infty} f_\alpha(z) \log \frac{\varphi_\alpha(z)}{f_\alpha(z)} dz$$

i) Si $\frac{\varphi_\alpha(Z_\alpha)}{f_\alpha(Z_\alpha)} = E\left[\frac{\varphi_\alpha(Z_\alpha)}{f_\alpha(Z_\alpha)}\right] = 1$

alors $\varphi_\alpha(Z_\alpha) = f_\alpha(Z_\alpha)$ et donc quelque soit la valeur z prise par Z_α,

$$\varphi_\alpha(z) = f_\alpha(z)$$

d'où

$$\varphi_\alpha = f_\alpha$$

ii) Si $\frac{\varphi_\alpha(Z_\alpha)}{f_\alpha(Z_\alpha)} \neq E\left[\frac{\varphi_\alpha(Z_\alpha)}{f_\alpha(Z_\alpha)}\right] = 1$

Posons $\frac{\varphi_\alpha(Z_\alpha)}{f_\alpha(Z_\alpha)} = U$ et $E\left[\frac{\varphi_\alpha(Z_\alpha)}{f_\alpha(Z_\alpha)}\right] = u$

par stricte convexité de la fonction $h = -\log$, on a :

$$h(U) > h(u) + \nabla h(u)(U - u)$$

ce qui donne l'inégalité de Jensen stricte :

$$E\left[h(U)\right] > h(u) = 0$$

d'où

$$\varphi_\alpha \neq f_\alpha \Longrightarrow I(\alpha) > 0 \quad \blacksquare$$

$I(\alpha)$ est donc minimum uniquement lorsque f_α est une densité normale.

existe si et seulement si toutes les structures marginales de dimension 1 associées sont normales.

Il suffit donc de montrer que la valeur maximum de $I(\alpha)$ est nulle pour conclure à une structure multidimensionnelle normale.

Dans l'optique de la théorie des tests, au vu de l'information dont on dispose, celle du tableau X, si l'on trouve une estimation $\hat{I}(\alpha)$ de $I(\alpha)$, on pourra procéder de la manière suivante :

Trouver la direction α_o qui maximise $\hat{I}(\alpha)$ et tester $I(\alpha_o) = 0$ pour conclure à une structure multidimensionnelle normale.

En maximisant $I(\alpha)$ par rapport à α on peut donc

- soit trouver une structure normale de Ω,
- soit trouver des structures projetées que l'on peut appeler "éloignées" de la structure normale :

Dans une certaine mesure, l'entropie (ou l'information) nous donne comme structure "intéressante" des structures projetées chaotiques de Ω à la manière de l'entropie d'un système physique qui "peut être considérée comme mesurant le désordre de ce système" (Rényi, 1966).

Remarque
========

Si l'on se restreint à une analyse sur un sous-ensemble borné

de $\mathbb{R}^n$ du type $[a,b]$ $(-\infty < a < b < +\infty)$ [illegible]

contre-indiqué vu que toutes les valeurs d'échantillon dont on dispose sont bornées, on peut alors considérer une entropie modifiée de la forme :

$$I'(\alpha) = -\int_a^b f_\alpha(z) \log \frac{\Psi_\alpha(z)}{f_\alpha(z)} \, dz$$

où Ψ_α est la densité de la loi uniforme sur [a,b] :

$$\Psi_\alpha \equiv \frac{1}{b-a} \qquad \text{sur } [a,b]$$

les considérations ci-dessus se rapportant à la loi normale peuvent maintenant se rapporter à la loi uniforme.

Pour justifier l'expression de I' il faut de plus supposer que

$$\text{si } z \notin [a,b], \quad f_\alpha(z) = 0$$

I' est borné inférieurement par 0 et cette borne n'est atteinte que pour $f_\alpha = \Psi_\alpha$ (même démonstration que celle de la proposition précédente). ■

On a abordé plus haut le problème de l'estimation de $I(\alpha)$ que l'on va traiter maintenant.

On va exposer brièvement une manière d'estimer $I(\alpha)$ sachant qu'on ne connaît pas f_α .

Il est possible évidemment de trouver une estimation de f_α dans l'échantillon $\{ z_1, \dots, z_n \}_\alpha$.

Cependant, soient κ_3 et κ_4 les cumulants de la distribution de Z_α .

On a l'approximation suivante :

Proposition

Lorsque $\mu = E(Z_\alpha) = 0$ et $\operatorname{Var} Z_\alpha = 1$

$$I(\alpha) \cong [\, \kappa_3^2 + \tfrac{1}{4} \kappa_4^2 \,]\ \tfrac{1}{12}$$

en effet :

posons $$\varepsilon_\alpha(z) = \frac{f_\alpha(z) - \varphi_\alpha(z)}{\varphi_\alpha(z)}$$

On a les propriétés P_1, P_2, P_3 :

$$(P_1) \quad \int_{-\infty}^{+\infty} \varphi_\alpha(z)\varepsilon_\alpha(z)\, dz = 0$$

(car f_α et φ_α sont deux densités sur $(-\infty, +\infty)$)

$$(P_2) \quad \int_{-\infty} \varphi_\alpha(z)\varepsilon_\alpha(z)\, z\, dz = 0$$

(car $\mu = E(Z_\alpha)$)

$$(P_3) \quad \int_{-\infty}^{+\infty} \varphi_\alpha(z)\varepsilon_\alpha(z)\, z^2\, dz = 0$$

(car la variable aléatoire de densité φ_α a même variance et même espérance que Z_α).

Ainsi, lorsque $\varepsilon(z)$ est "assez petit", $f_\alpha(z)\log f_\alpha(z)$ admet l'approximation :

$$\varphi_\alpha(z)(1 + \varepsilon_\alpha(z))(\log \varphi_\alpha(z) + \varepsilon_\alpha(z) - \tfrac{1}{2} \varepsilon_\alpha^2(z))$$

ou encore :

$$\varphi_\alpha(z)\varepsilon_\alpha(z) + \tfrac{1}{2} \varphi_\alpha(z)\varepsilon_\alpha^2(z) + \varphi_\alpha(z)\varepsilon_\alpha(z)\log \varphi_\alpha(z) + \varphi_\alpha(z)\log \varphi_\alpha(z)$$

et sous de bonnes conditions d'intégrabilité, les propriétés P_1, P_2 et P_3 nous donnent :

i) $$\int_{-\infty}^{+\infty} \varphi_\alpha(z)\varepsilon_\alpha(z)dz = \int_{-\infty}^{+\infty} \varphi_\alpha(z)\varepsilon_\alpha(z)\log \varphi_\alpha(z)dz = 0$$

ii) $$\int_{-\infty}^{+\infty} \varphi_\alpha(z)\log \varphi_\alpha(z)dz = -\log \sqrt{2\pi e}$$

iii) $I(\alpha) \cong \frac{1}{2} \int_{-\infty}^{+\infty} \varphi_\alpha(z) \varepsilon_\alpha^2(z) dz$

En développant f_α en série de Gram-Charlier (Kendall et Stuart, 1976) jusqu'à l'ordre 4, on peut écrire du fait que φ_α est la densité d'une loi normale centreé réduite :

$$f_\alpha(z) = \varphi_\alpha(z) \left(1 + \frac{\kappa_3}{6} H_3(z) + \frac{\kappa_4}{24} H_4(z)\right)$$

avec $$\int_{-\infty}^{+\infty} H_m(z) H_{m'}(z) \varphi_\alpha(z) dz = \begin{cases} 0 & \text{si } m \neq m' \\ m! & \text{si } m = m' \end{cases}$$

de sorte qu'en identifiant $\varepsilon_\alpha(z)$ et $\frac{\kappa_3}{6} H_3(z) + \frac{\kappa_4}{24} H_4(z)$, il vient :

$$I(\alpha) \cong [\kappa_3^2 + \tfrac{1}{4} \kappa_4^2] \tfrac{1}{12} \quad \blacksquare$$

A partir d'une telle approximation et d'une estimation $\hat{\kappa}_3^2$ (resp. $\hat{\kappa}_4^2$) de κ_3^2 (resp. κ_4^2), on a une estimation de $I(\alpha)$:

$$\hat{I}(\alpha) = [\hat{\kappa}_3^2 + \tfrac{1}{4} \hat{\kappa}_4^2] \tfrac{1}{12}$$

On prend :

$$\hat{\kappa}_3^2 = [\tfrac{1}{n} \sum_1^n z_i^3]^2$$

$$\hat{\kappa}_4^2 = [\tfrac{1}{n} \sum_1^n z_i^4 - 3]^2$$

Remarque
========

L'approximation énoncée dans la proposition ci-dessus n'a lieu que pour un bon comportement de ε.

Cependant, elle justifie l'utilisation d'un indice de projection basé sur les cumulants d'ordre supérieur ou égal à 3 de f_α du fait :

- de la facilité de calcul de ces cumulants dans l'échantillon des points projetés

- de la nullité des cumulants d'ordre supérieur ou égal à 3 d'une distribution normale : un indice de projection à valeurs positives tel que celui que l'on a trouvé est alors minimum dans le cas d'une distribution normale.

Mentionnons enfin que κ_3 et κ_4 interviennent dans la définition des paramètres de forme ("skewness" et "kurtosis") utilisés dans l'étude du "degré de non-normalité d'une distribution" (Guiard, 1984). ■

CHAPITRE 2

Echelles, Encadrements et Invariance

Les exemples qui ont été présentés précédemment ont permis d'aborder des questions d'échelle concernant les données brutes constituant le tableau X ainsi que des questions concernant le sous-ensemble d'observation des points de Ω.

2.1 ECHELLES

L'utilité de transformer chaque échantillon $\underline{x}^i$ par homothétie peut apparaître soit dans la construction de tests statistiques relativement simples, soit dans l'interprétation des résultats obtenus, soit enfin dans la représentation graphique du nuage.

2.1.1 Changement d'échelle pour la construction d'un test

Dans le modèle de régression de l'exemple a), on s'intéresse non seulement à la dépendance linéaire des y_i par rapport aux x_{ij}, $i = 2, \ldots, p$, mais on veut aussi apprécier le caractère

explicatif des x_{ij}, $i = 2, \ldots, p$ par rapport aux y_j ce qui se traduit par :

$$\beta_i \neq 0 \quad , i = 2, \ldots, p$$

(en effet, $\beta_i = 0$ n'exclut pas la dépendance linéaire mais exclut le caractère explicatif).

On note Y la 1ère variable et y, une réalisation de Y.

On note $X_2, \ldots, X_p$ respectivement les $2^{\text{ième}}, \ldots, p^{\text{ième}}$ variables et x_i une réalisation de X_i.

On considère la matrice diagonale W des variances empiriques de Y, $X_2, \ldots, X_p$:

$$W = \begin{pmatrix} \frac{1}{n}\sum_1^n (y_j-\bar{y})^2 & & & \\ & \frac{1}{n}\sum_j (x_{2j}-\bar{x}^2)^2 & & \\ & & \ddots & \\ & & & \frac{1}{n}\sum_j (x_{pj}-\bar{x}^p)^2 \end{pmatrix}$$

La variable discriminante t du test de signification de Student sur chaque β_i pris individuellement est :

$$t = \frac{\rho_{yx_i \cdot x_2 \ldots x_{i-1} x_{i+1} \ldots x_p}}{\sqrt{1 - \rho^2_{yx_i \cdot x_2 \ldots x_{i-1} x_{i+1} \ldots x_p}}}$$

où $\rho_{yx_i.x_2\cdots x_{i-1}x_{i+1}\cdots x_p}$ est le coefficient de corrélation partielle empirique entre Y et X_i.

On constate que t est calculée sur un tableau de données centré, réduit puisque $\rho_{yx_i.x_2\cdots x_{i-1}x_{i+1}\cdots x_p}$ est une fonction des corrélations Corr (Y,X_i) et Corr (X_i,X_j) apparaissant dans le tableau :

$$W^{-\frac{1}{2}} X \left(I - \frac{SS'}{n}\right) X' W^{-\frac{1}{2}}$$

le changement d'échelle envisagé pour le test correspond à la transformation sur les données :

$$X \longmapsto W^{-\frac{1}{2}} X$$

On sait que des changements d'échelle de ce type trouvent leur justification dans les théorèmes limites du calcul des probabilités.

2.1.2 Changement d'échelle pour l'interprétation des résultats

Rapportons-nous ici à l'exemple c) concernant l'analyse factorielle discriminante.

Comme on l'a vu, l'application de l'analyse en composantes principales au tableau de données $\tilde{X}$ du nuage des centres de gravité des classes, conduit à la transformation sur les données :

$$\tilde{X} \longmapsto \sqrt{n}\ \left(\tilde{X}\left(I - \frac{SS'}{\tilde{n}}\right)\tilde{X}'\right)^{-\frac{1}{2}} \tilde{X}$$

Une telle transformation permet d'interpréter le type de projection du nuage de ces centres de gravité comme celle qui optimise à la fois une distance intraclasse et une distance interclasse calculées sur Ω.

Cette transformation a la propriété suivante : la matrice de covariance empirique transformée est égale à la matrice identité p x p.

Ce n'est pas la seule transformation qui ait cette propriété.

Ceci nous amène à définir l'équivalence de deux transformations Q_1 et Q_2 sur un tableau de données X :

Définition

$X \longmapsto Q_1X$ et $X \longmapsto Q_2X$
sont deux transformations équivalentes si :

$$\frac{1}{n} Q_1X\left(I - \frac{SS'}{n}\right)X'Q_1' = \frac{1}{n} Q_2X\left(I - \frac{SS'}{n}\right)X'Q_2' = I \quad \blacksquare$$

On pose :

$$V = \frac{1}{n} X \left(I - \frac{SS'}{n}\right) X'$$

Proposition

Si V est inversible, la transformation $X \longmapsto V^{-1/2} X$ est équivalente à la transformation $X \longmapsto DHX$ où :

- H est une matrice de changement d'axes orthogonale
- et D est une matrice de changement d'échelle sur chacun des nouveaux axes précités.

En effet :

Soit $H'D^{-2}H$ une décomposition spectrale de la matrice symétrique définie positive V (H, orthogonale).

On a donc :

$$V^{-1/2} = H'DH$$

et :

$$V^{-1/2} \, V \, V^{-1/2} = I = DHH'D^{-2}HH'D = DHVH'D \quad \blacksquare$$

Remarque

Il n'est pas possible en général de réaliser sur X d'abord un changement d'échelle et ensuite un changement d'axes de manière à aboutir à une transformation équivalente à $V^{-1/2}$.

En effet, supposons qu'il existe Δ, diagonale, inversible, et T, orthogonale, telles que :

$$T\Delta H'D^{-2}H\Delta T' = I$$

$$H'D^{-2}H = \Delta^{-2}$$

$$D^{-2}H = H\Delta^{-2}$$

les colonnes de H sont donc des vecteurs propres pour D^{-2} qui est diagonale.

La décomposition de l'espace $\mathbb{R}^p$ en sous-espaces propres, pour la matrice D^{-2}, est unique lorsque chaque sous-espace correspond à une seule valeur propre de D^{-2} de multiplicité supérieure ou égale à 1. A cette décomposition, on devrait associer une matrice H de vecteurs propres, nécessairement bloc-diagonale avec au moins 2 blocs lorsque D^{-2} n'est pas scalaire. H n'a généralement pas cette forme. $\blacksquare$

En conclusion, la méthode d'analyse du tableau $\tilde{X}$ décrite dans l'exemple c) tient compte d'un changement d'échelle sur le tableau $\tilde{X}$ exprimé dans un nouveau repère orthonormé.

Lorsqu'on leur applique la transformation que l'on vient d'étudier, les données sont dites "sphérées" : la matrice de covariance empirique de données sphérées est l'identité par définition.

2.1.3 Changement d'échelle pour la représentation graphique

Ainsi qu'on l'a déjà indiqué, on pourra être amené à rechercher certaines structures homogènes du nuage Ω et à les faire

En analyse en composante principale, par exemple, c'est une structure globale homogène que l'on recherche puisque l'on minimise une inertie totale.

L'hyperplan de projection dépend de tout changement d'échelle opéré sur chacun des axes initiaux.

La visualisation, par projection, de telles structures homogènes est donc liée à l'échelle définie sur chaque axe.

Il y a un certain arbitraire à définir les échelles lorsque l'on travaille sur des variables dites "incommensurables".

Pour pallier cette difficulté, on a recours le plus souvent à la standardisation des variables : le changement d'échelle sur chaque axe est celui qui correspond à la transformation sur les données déjà rencontrée :

$$X \longmapsto W^{-1/2}X$$

Autrement dit, chaque élément d'échantillon d'une variable est "mesuré en unité d'écart-type" (Chandon et Pinson, 1981), lequel est calculé sur l'ensemble des éléments de l'échantillon.

Ainsi ces mesures sont invariantes par changement d'échelle, c'est-à-dire que si Δ est une matrice diagonale (définie positive),

$$W^{-1/2}\, X \;=\; (\Delta^2\, W)^{-1/2}\, \Delta\, X$$

déformations dans le nuage : ainsi, un exemple rapporté par Chandon et Pinson (1981) prouve qu'une telle standardisation " ne préserve pas l'ordre des distances euclidiennes obtenues sur les données brutes", ces distances étant, dans notre cas, celles qui existent entre les points de Ω.

2.2 ENCADREMENT

Le problème envisagé est le suivant : étant donné une variable Z_α dont on connaît un échantillon sous forme vectorielle $\alpha'X$, quel est son support ?

En supposant qu'il s'agit d'une variable aléatoire, chaque fois que l'on fait une hypothèse sur sa distribution, on détermine, de facto, son support.

Soit $z \longmapsto F_\alpha(z)$, la fonction de répartition de Z_α et $z \longmapsto F'_\alpha(z)$, sa fonction de répartition empirique.

Si le support de Z_α est un intervalle Θ de $\mathbb{R}$, on a la propriété $\mathbb{P}_o$:

$$\forall\, z \notin \Theta\ ,\ \forall\, z' \in \Theta,$$

$$z < z' \Longrightarrow F_\alpha(z) = 0$$
$$z \geq z' \Longrightarrow F_\alpha(z) = 1$$

(par convention, si $\Theta = (-\infty,+\infty)$, $z = \pm\infty$).

$\forall\, z \notin \hat{\Theta}\ ,\ \forall\, i \in (1, \ldots, n)$

$$z < z_i \Longrightarrow F'_\alpha(z) = 0$$
$$z \geq z_i \Longrightarrow F'_\alpha(z) = 1$$

est une estimation du support de Z_α.

On voit que, si $\hat{\Theta} = [\hat{a}, \hat{b}]$ est une estimation du support de Z_α, $\hat{\theta}$ est tel que

$$[\hat{a}, \hat{b}] \supset [z_{(1)}, z_{(n)}]$$

On peut s'intéresser, du fait de la forme de cette contrainte, à estimer a et b de manière à ne tenir compte que de $z_{(1)}$ et $z_{(n)}$ et non des autres statistiques d'ordre $z_{(2)}, \ldots, z_{(n-1)}$.

Ce faisant, il est possible de trouver un intervalle $\hat{\Theta}$ qui soit ou bien dépendant de α ou bien indépendant de α.

2.2.1 Une manière de calculer $\hat{\Theta}$ pour α donné

On cherche un intervalle $\hat{\Theta}$ borné ce qui se justifie par le fait que les valeurs d'échantillon sont bornées.

En supposant que Z_α possède une distribution uniforme, les estimateurs du maximum de vraisemblance $\hat{a}_o$ et $\hat{b}_o$ de a et b sont des L-estimateurs (Lecoutre et Tassi, 1987) ne dépendant que de $z_{(1)}$ et $z_{(n)}$:

$$\hat{b}_o = z_{(n)}$$

les L-estimateurs non biaisés $\hat{a}_1$ et $\hat{b}_1$ obtenus par combinaisons linéaires des précédents sont :

$$\hat{a}_1 = z_{(1)} - \frac{1}{n-1}\,(z_{(n)} - z_{(1)})$$
$$\hat{b}_1 = z_{(n)} + \frac{1}{n-1}\,(z_{(n)} - z_{(1)})$$

Lorsque, comme dit Friedman (1987), "on est intéressé par toute distribution considérée comme alternative à la distribution uniforme" (en fait Friedman parle de la distribution normale), ces estimateurs de a et b sont tout-à-fait opportuns pour estimer Θ.

Considérons maintenant, pour tout couple de réels (r,s), la propriété :

$$\mathbb{P}(r, s) \ : \begin{cases} z_{(1)} - r & < \dfrac{s - r}{n} \\ s - z_{(n)} & \leq \dfrac{s - r}{n} \end{cases}$$

On remarque que $\mathbb{P}(\hat{a}_1, \hat{b}_1)$ est vérifiée puisqu'en effet :

$$\hat{b}_1 - \hat{a}_1 = \frac{n+1}{n-1}\,(z_{(n)} - z_{(1)})$$

et par conséquent,

$$z_{(1)} - \hat{a}_1 = \frac{\hat{b}_1 - \hat{a}_1}{n+1} < \frac{\hat{b}_1 - \hat{a}_1}{n}$$
$$\hat{b}_1 - z_{(n)} = \frac{\hat{b}_1 - \hat{a}_1}{n+1} \leq \frac{\hat{b}_1 - \hat{a}_1}{n}$$

Remarque 1
==========

Etant donné une subdivision de $[\hat{a}, \hat{b}[$ en n sous-intervalles I_u, $u = 1, \ldots, n$ semi-ouverts à droite, contigus, de même diamètre, si chaque I_u contient un point de $(z_1, \ldots, z_n)_\alpha$ alors $\mathbb{P}(\hat{a}, \hat{b})$ est vérifiée. ■

Remarque 2
==========

L'estimation du support de Z_α n'apporte pas nécessairement d'information supplémentaire dans l'analyse exploratoire de données, dans la mesure, par exemple, où l'indice de projection ne dépend de Z_α qu'à travers les valeurs d'échantillon constituant le vecteur $\underline{Z}_\alpha$, ce qui est le cas entre autre dans l'exemple b) :

$$\tilde{I} = \alpha' X (I - \frac{SS'}{n}) X' \alpha$$

$$= \underline{Z}_\alpha (I - \frac{SS'}{n}) \underline{Z}'_\alpha$$

(α inclu le facteur matriciel $M^{1/2}$). ■

2.2.2 Une manière de calculer $\hat{\Theta}$ indépendamment de α

A ce stade, on est contraint à préciser quelle direction de projection on utilise par rapport à un sous-espace de projection donné.

Pour fixer les idées, on adopte la projection orthogonale.

une projection sur un espace à une dimension

donne le nuage projeté noté vectoriellement $\alpha' X$ $(\alpha, p \times 1)$.

Prendre la projection orthogonale, c'est imposer $\alpha'\alpha = 1$.

Les solutions du problème :

$$\begin{cases} \text{Opt } \alpha' X \\ \alpha'\alpha = 1 \end{cases}$$

conduisent à un encadrement des composantes de $\alpha' X$, exprimé à l'aide de la forme quadratique $\| \ \|^2$:

$$\forall j , \qquad - \|\underline{x}_j\| \leq \alpha' \underline{x}_j \leq \|\underline{x}_j\|$$

En prenant

$$\hat{\Theta} = [-\max \|\underline{x}_j\| , \max \|\underline{x}_j\|]$$

on détermine, pour Z_α , un support estimé indépendant de α.

Remarque
========

Lorsque X est un tableau centré, réduit, on peut simplifier le résultat précédent. En effet, on a dans ce cas :

$$\|\underline{x}^i\| = \sqrt{n} \implies \forall i,j \quad |x_{ij}| \leq \sqrt{n} \implies$$

$$\forall j , \quad \|\underline{x}_j\| \leq \sqrt{np}$$

$$\hat{\theta} = [\ -\sqrt{np}\ ,\ \sqrt{np}\ ] \qquad \blacksquare$$

On voit que l'étude du nuage Ω, en l'absence de toute perturbation (X ne subit pas de transformation), peut s'effectuer sur le domaine borné de $\mathbb{R}^p$ (rectangle) :

$$RC = [\ -\max \|\underline{x}_j\| \ ,\ \max \|\underline{x}_j\| \ ]^p$$

2.3 INVARIANCE

On s'intéresse à l'invariance des structures de Ω lorsqu'on opère certaines transformations, soit du tableau X, soit de l'échantillon α'X des données projetées.

2.3.1 Transformation du tableau X

Toute transformation du tableau X, c'est-à-dire des données brutes, peut entraîner des modifications sur l'échantillon $(\ z_1, \ldots, z_n)_\alpha$.

Il est un fait, par exemple, qu'un changement d'échelle du type $X \longmapsto \Delta X$ (Δ, diagonale, définie positive) ne se convertit en une homothétie sur l'échantillon que dans la direction des vecteurs de base.

En revanche, une translation sur les points de Ω induit une translation sur les éléments de $(\ z_1, \ldots, z_n\ )_\alpha$.

Ω translatés par $\underline{t}$ est :

$$X_t = X + \underline{t}s'$$

On constate qu'effectuer $\alpha'X_t$ revient à ajouter à chaque z_i la même quantité réelle $\alpha'\underline{t}$.

Ainsi, comme on devait s'y attendre, dans un problème d'exploration par indice de projection, tout changement d'origine de l'espace ne peut pas affecter l'étude des structures contenues dans le nuage.

Pour mémoire, on peut signaler qu'on a déjà translaté X lorsqu'on a utilisé le centrage :

$$X \longmapsto X(I - \frac{SS'}{n})$$

2.3.2 Transformation de l'échantillon α'X

Comme on vient de le voir, certaines transformations de Ω peuvent laisser invariante la détermination de ses structures déduites de l'analyse de α'X.

Il faut voir maintenant dans quelle mesure une transformation de α'X entraîne une transformation de l'indice de projection.

Une telle transformation n'a évidemment d'intérêt que dans la mesure où elle peut apporter des simplifications dans le calcul de

A ce sujet, on peut définir trois classes d'indice (Huber,1985) :

Pour cela, précisons que tout indice K peut-être considéré soit comme une fonction de α soit comme une fonction de $\underline{Z}_\alpha$ (nx1).

Etant donné deux constantes réelles λ_1 et λ_2 ($\lambda_1 \neq 0$), on a les définitions suivantes :

Définition 1

K est linéaire (de classe 1) si

$$K(\underline{Z}_\alpha\lambda_1 + S\lambda_2) = \lambda_1 K(\underline{Z}_\alpha) + \lambda_2 \quad \blacksquare$$

Définition 2

K est invariant par translation et "équivariant" par changement d'échelle (de classe 2) si

$$K(\underline{Z}_\alpha\lambda_1 + S\lambda_2) = |\lambda_1|\, K(\underline{Z}_\alpha) \quad \blacksquare$$

Définition 3

K est invariant par affinité (de classe 3) si

$$K(\underline{Z}_\alpha\lambda_1 + S\lambda_2) = K(\underline{Z}_\alpha) \quad \blacksquare$$

Si K appartient à la classe 2 ou à la classe 3, une transformation affine de $\underline{Z}_\alpha$ ne modifiera pas les solutions éventuelles d'un

Remarque

Dans l'exemple d), l'indice I(.) appartient à la classe 3.

En effet, en notant g_α la densité de $\lambda_1 Z_\alpha + \lambda_2$, on a :

$$g_\alpha(y) = \frac{1}{|\lambda_1|} f_\alpha\left(\frac{y-\lambda_2}{\lambda_1}\right)$$

d'où :

$$I(\alpha) = -\int f_\alpha(z) \log \frac{\varphi_\alpha(z)}{f_\alpha(z)}\, dz$$

$$= -\int g_\alpha(y) \log \frac{\lambda_1}{\lambda_1} \frac{\varphi_\alpha\left(\frac{y-\lambda_2}{\lambda_1}\right)}{f_\alpha\left(\frac{y-\lambda_2}{\lambda_1}\right)}\, dy$$

$$= -\int g_\alpha(y) \log \frac{\varphi'_\alpha(y)}{g_\alpha(y)}\, dy$$

où φ'_α est la densité d'une variable aléatoire normale de même espérance et de même variance que $\lambda_1 Z_\alpha + \lambda_2$. $\blacksquare$

CHAPITRE 3

Le Coefficient de Concentration; Généralisation

3.1 LE COEFFICIENT DE CONCENTRATION

On présentera par la suite un nouvel indice de projection construit d'une manière qui ne tienne pas compte de la distance interpoints.

Il est clair, par exemple, que, si l'on ne pondère pas les points de Ω, la détermination d'un hyperplan factoriel à l'aide de la minimisation d'une inertie $\tilde{I} = \sum_{j=1}^{n} \|\underline{x}_j - \hat{\underline{x}}_j\|^2$, prend en compte, de façon primordiale, les points éloignés de cet hyperplan; c'est-à-dire les points $\underline{x}_j$ tels que le terme $\|\underline{x}_j - \hat{\underline{x}}_j\|^2$ ait une forte contribution dans le calcul de $\tilde{I}$.

Pour se dégager de la notion de distance interpoints, le coefficient de concentration considéré plus tard comme un indice de projection, présentera un intérêt essentiel.

Le coefficient de concentration a été introduit par C. Tricot (1971) pour analyser certaines structures d'ensembles bornés sur $\mathbb{R}$. Comme il le montre, c'est un nouvel indice pour l'étude de la concentration de populations sur un territoire en géographie.

Ce coefficient est défini comme une application C de $\mathbb{R}^n$ dans $\mathbb{R}$ de la façon suivante :

Soit une suite finie $(y_i)_{i=1,\ldots,n}$ de points de $\mathbb{R}$ $(n \geq 1)$.

On notera, à partir de maintenant, $\underline{Y}$, le vecteur $(y_1, \ldots, y_n)'$ et $\{y_1, \ldots, y_n\}$ le sous-ensemble de $\mathbb{R}$ des éléments distincts de la suite (y_i).

Soit l'intervalle défini pour ε strictement positif quelconque :

$$I_y^{\varepsilon} = [\, y_{(1)} \, , \, y_{(n)} + \varepsilon \,[$$

Soit une partition de I_y^{ε} en n^m intervalles $(m \geq 0)$, contigus, de même diamètre :

$$I_{u,m}^{\varepsilon} = \left[\, y_{(1)} + \frac{u-1}{n^m}(y_{(n)} - y_{(1)} + \varepsilon) \, , \, y_{(1)} + \frac{u}{n^m}(y_{(n)} - y_{(1)} + \varepsilon) \,\right[$$

$1 \leq u \leq n^m$

La suite finie $(I^{\varepsilon}_{u,m})_{u=1,\ldots,n^m}$ est une grille G_m de niveau m sur I^{ε}_{y}.

$I^{\varepsilon}_{u,m}$ est une maille-m de la grille G_m.

(G_m) est un réseau sur I^{ε}_{y} .

Pour les différentes définitions et références au sujet de ces notions de maille, grille, et réseau, on consultera l'annexe 1 .

On remarque que le nombre de mailles-m est égal à n^m.

On pose :

i) $$K^{\varepsilon}_{u,m}(\underline{Y}) = \sum_{i=1}^{n} 1_{I^{\varepsilon}_{u,m}}(y_i)$$

$K^{\varepsilon}_{u,m}(\underline{Y})$ est le nombre de y_i appartenant à la maille-m $I^{\varepsilon}_{u,m}$.

ii) $$N^{\varepsilon}_{u,m}(\underline{Y}) = n - \sum_{v=(u-1)n+1}^{un} 1_{\mathbb{N}^*}\left(\sum_{i=1}^{n} 1_{I^{\varepsilon}_{v,m+1}}(y_i) \right)$$

$N^{\varepsilon}_{u,m}(\underline{Y})$ est le nombre de mailles-(m+1) inclues dans la maille-m $I^{\varepsilon}_{u,m}$ et ne contenant pas de y_i.

On note que, quelque soit u, m et ε,

$K^{\varepsilon}_{1,0}(\underline{Y}) = n$

$0 \le N^{\varepsilon}_{1,0}(\underline{Y}) \le n-1$

$0 \le N^{\varepsilon}_{u,m}(\underline{Y}) \le n$

$$C(\underline{Y}) = \sum_{m=0}^{\infty} \frac{1}{n^{m+1}} \lim_{\varepsilon \to 0} \sum_{u=1}^{n^m} \frac{K^{\varepsilon}_{u,m}(\underline{Y})}{n} N^{\varepsilon}_{u,m}(\underline{Y}) \quad \blacksquare$$

En posant $n_{m+1} = \lim_{\varepsilon \to 0} \sum_{u=1}^{n^m} \frac{K^{\varepsilon}_{u,m}(\underline{Y})}{n} N^{\varepsilon}_{u,m}(\underline{Y})$

on a :

$$C(\underline{Y}) = \sum_{1}^{\infty} \frac{n_m}{n^m}$$

Pour les développements qui vont suivre, on a besoin de considérer les fonctions suivantes, définies sur $\mathbb{R}^{+*}$(u,m,fixés) :

$\varphi_1 : \varepsilon \longmapsto N^{\varepsilon}_{u,m}(\underline{Y})$

$\varphi_2 : \varepsilon \longmapsto K^{\varepsilon}_{u,m}(\underline{Y})$

$\varphi_3 : \varepsilon \longmapsto \text{Card}\,(I^{\varepsilon}_{u,m} \cap \{y_1, \ldots, y_n\})$

Lemme 1

Les fonctions φ_i , i = 1, 2, 3, sont des fonctions en escaliers. ■

Lemme 2

$\forall \varepsilon' \ge 0$, $\exists \varepsilon_o > 0$ tel que les trois fonctions φ_i , i=1,2,3 soient constantes sur $]\varepsilon', \varepsilon' + \varepsilon_o[$. ■

Remarque

Les φ_i , i = 1, 2, 3, sont continues à gauche.

sont semi-ouverts à droite. ■

Lemme 3

On a les différents résultats suivants :

a) $\forall u,m,\varepsilon,\quad K^{\varepsilon}_{u,m}(\underline{Y}) \neq 0 \Longrightarrow 0 \leq \frac{N^{\varepsilon}_{u,m}(\underline{Y})}{n} \leq \frac{n-1}{n}$

b) $\forall m,\varepsilon,\quad \sum_{u=1}^{n^m} K^{\varepsilon}_{u,m}(\underline{Y}) = n$

c) Si $\forall i,j,\ Y_i = Y_j$ alors $\forall m,\ n_m = n-1$

d) $\forall M \geq 1\ ,\ \sum_{m=M}^{\infty} \frac{n_m}{n^m} \leq \frac{1}{n^{M-1}} \leq 1$

Le point d) résulte du fait que, d'après a) et b),

$$\forall m \geq 0,\ n_{m+1} \leq \lim_{\varepsilon \to 0} \frac{n-1}{n} \sum_{u=1}^{n^m} K^{\varepsilon}_{u,m}(\underline{Y}) = n-1$$

Ainsi, lorsque $M \geq 1$,

$$\sum_{m=M-1}^{\infty} \frac{n_{m+1}}{n^{m+1}} \leq \frac{n-1}{n^M} \frac{n}{n-1} = \frac{1}{n^{M-1}} \cdot \quad ■$$

Propriétés

1) $C(\underline{Y})$ est la somme d'une série convergente.
(voir d) dans le lemme 3)

En effet, le résultat d) du lemme 3, pour M = 1, implique $C(\underline{Y}) \leq 1$.

D'après le point c) , si $\forall i,j,\ Y_i = Y_j$, on a :

$$C(\underline{Y}) = 1$$

Réciproquement, supposons qu'il existe un couple (i,j) tel que $Y_i \neq Y_j$.

On a deux cas :

i) $n_1 \leq n-2$

D'après le résultat d) du lemme 3, pour M = 2,

$C(\underline{Y}) \leq \frac{n-2}{n} + \frac{1}{n} < 1$

ii) $n_1 = n-1$

D'après le lemme 2, on a :

$(R_1)\ \exists u_o,\ \exists \varepsilon_o>0 \mid \forall \varepsilon,\ 0<\varepsilon<\varepsilon_o\ ,\ \mathrm{Card}(I^{\varepsilon}_{u_o,1} \cap (Y_1,\ldots,Y_n))=n$

D'autre part, $\exists m_o \mid \forall \varepsilon,\ 0<\varepsilon<\varepsilon_o$

$$\frac{Y_{(n)}-Y_{(1)}+\varepsilon}{n^{m_o}} < \mid Y_i - Y_j \mid$$

et par conséquent, on a :

$(R_2)\ \forall \varepsilon,\ 0<\varepsilon<\varepsilon_o \quad Y_i \in I^{\varepsilon}_{u,m_o} \Longrightarrow Y_j \notin I^{\varepsilon}_{u,m_o}$

D'après (R_1) et (R_2), il existe un entier m_1 tel que

$m_1 = \mathrm{Max}(m \in \{1,\ldots,m_o-1\}; \exists u \text{ et } \exists \varepsilon'_o \leq \varepsilon_o \mid \forall \varepsilon, 0<\varepsilon<\varepsilon'_o\ , I^{\varepsilon}_{u,m} \ni Y_i \text{ et } I^{\varepsilon}_{u,m} \ni Y_j)$

Il existe donc u_1 tel que l'on ait, $\forall \varepsilon,\ 0<\varepsilon<\varepsilon'_o$,

$$\mathrm{Card}(I^{\varepsilon}_{u_1,m_1} \cap (Y_1,\ldots,Y_n)) \geq 2$$

$$K^{\varepsilon}_{u_1,m_1}(\underline{Y}) \geq 2$$

Et, par définition de m_1,

$$\forall \varepsilon,\ 0<\varepsilon<\varepsilon'_o\ ,\ N^{\varepsilon}_{u_1,m_1}(\underline{Y}) \leq n-2$$

D'où :

$$n_{m_1+1} \leq \lim_{\substack{\varepsilon\to 0\\ \varepsilon<\varepsilon'_o}} \left\{ (n-1)\sum_{u\neq u_1} \frac{K^{\varepsilon}_{u,m_1}(\underline{Y})}{n} + (n-2)\ \frac{K^{\varepsilon}_{u_1,m_1}(\underline{Y})}{n} \right\}$$

$$\leq \lim_{\substack{\varepsilon\to 0\\ \varepsilon<\varepsilon'_o}} \left\{ (n-1)\sum \frac{K^{\varepsilon}_{u,m_1}(\underline{Y})}{n} - \frac{K^{\varepsilon}_{u_1,m_1}(\underline{Y})}{n} \right\}$$

$$\leq n-1-\frac{2}{n}$$

$$< n-1$$

Ce qui implique :

$$\sum_1^{\infty} \frac{n_m}{n^m} < (n-1)\sum_1^{\infty} \frac{1}{n^m} = 1\ .$$

3) $C(\underline{Y}) \geq \frac{1}{n}$ et $C(\underline{Y}) = \frac{1}{n} \Longleftrightarrow n_1 = 0$

En effet,

i) si $n_1 \geq 1$, $C(\underline{Y}) > \frac{n_1}{n} \geq \frac{1}{n}$

ii) si $n_1 = 0$, on a d'après le lemme 2 les équivalences suivantes :

$$\exists \varepsilon_o \mid \forall \varepsilon,\ 0<\varepsilon<\varepsilon_o\ ,\ \forall u,\ \mathrm{Card}(I^{\varepsilon}_{u,o} \cap (y_1,\ldots,y_n))=1$$

$$\Longleftrightarrow$$

$$\exists \varepsilon_o \mid \forall \varepsilon,\ 0<\varepsilon<\varepsilon_o\ ,\ \forall u \begin{cases} m\geq 1 \Rightarrow \mathrm{Card}(I^{\varepsilon}_{u,m} \cap (y_1,\ldots,y_n)) \leq 1 \\ m=0 \Rightarrow \mathrm{Card}(I^{\varepsilon}_{u,o} \cap (y_1,\ldots,y_n)) = 1 \end{cases}$$

$$\Longleftrightarrow$$

$$\begin{cases} \forall m \geq 1\ ,\ n_{m+1} = n-1 \\ n_1 = 0 \end{cases}$$

4) $C(\underline{Y})$ est rationnel et $\exists m' \mid \forall m \geq m'\ ,\ n_m = n_{m'}$

En effet,

$\exists \varepsilon_o\ ,\ \exists m_o \mid \forall \varepsilon\ ,\ 0<\varepsilon<\varepsilon_o\ ,\ \forall i,j\ ,$

$$y_i \neq y_j \Longrightarrow \frac{y_{(n)}-y_{(1)}+\varepsilon}{n^{m_o}} < |\ y_i - y_j\ |$$

D'où,

$$m \geq m_o \Longrightarrow n_{m+1} = n-1$$

Ainsi,

$$C(\underline{Y}) = \sum_1^{m_o} \frac{n_m}{n^m} + (n-1)\sum_{m_o+1}^{\infty} \frac{1}{n^m}$$

$$= \sum_1^{m_o} \frac{n_m}{n^m} + \frac{1}{n^{m_o}} \in \mathbb{Q}$$

De plus, $\forall m \geq m_o + 1$, en posant $m' = m_o + 1$,

$$n_m = n_{m'} = n-1\ .\ \blacksquare$$

$$K^{\varepsilon}_{u_1,m_1}(\underline{Y}) \geq 2$$

Et, par définition de m_1,

$$\forall \varepsilon,\ 0<\varepsilon<\varepsilon''_0\ ,\ K^{\varepsilon}_{u_1,m_1}(\underline{Y}) \leq m-2$$

D'où :

$$m_{m_1+1} \leq \lim_{\substack{\varepsilon\to 0\\ \varepsilon<\varepsilon''_0}} \left\{ (m-1)\sum_{u\neq u_1} \frac{K^{\varepsilon}_{u,m_1}(\underline{Y})}{m} + (m-2)\ \frac{K^{\varepsilon}_{u_1,m_1}(\underline{Y})}{m} \right\}$$

$$\leq \lim_{\substack{\varepsilon\to 0\\ \varepsilon<\varepsilon''_0}} \left\{ (m-1)\sum \frac{K^{\varepsilon}_{u,m_1}(\underline{Y})}{m} - \frac{K^{\varepsilon}_{u_1,m_1}(\underline{Y})}{m} \right\}$$

$$\leq m-1-\frac{2}{m}$$

$$< m-1$$

Ce qui implique :

$$\sum_1^{\infty} \frac{m_m}{m^m} < (m-1)\sum_1^{\infty} \frac{1}{m^m} = 1 \quad .$$

3) $C(\underline{Y}) \geq \frac{1}{m}$ et $C(\underline{Y}) = \frac{1}{m} \Longleftrightarrow m_1 = 0$

En effet,

i) si $m_1 \geq 1$, $C(\underline{Y}) > \frac{m_1}{m} \geq \frac{1}{m}$

ii) si $m_1 = 0$, on a d'après le lemme 2 les équivalences suivantes :

Donc $Y_{(1)} \neq Y_{(n)}$.

On a, d'après le lemme 2 :

$$\exists \varepsilon_o \,|\, \forall \varepsilon, 0<\varepsilon<\varepsilon_o\ , n_{m+1} = \sum_{u=1}^{n^m} \frac{K^{\varepsilon}_{u,m}(\underline{Y})}{n} N^{\varepsilon}_{u,m}(\underline{Y}) \geq \sum_{u=1}^{n^m} \frac{K^{\varepsilon}_{u,m}(\underline{Y})}{n} (n-K^{\varepsilon}_{u,m}(\underline{Y}))$$

Ce minorant de n_{m+1} se réécrit, d'après le résultat b) du lemme 3 :

$$n - \frac{1}{n}\sum_{u=1}^{n^m} (K^{\varepsilon}_{u,m}(\underline{Y}))^2$$

Comme on envisage cette quantité pour $m+1\geq 2$, la somme comprend plus d'un terme.

De plus, au moins deux des termes de la somme sont non nuls : en effet, puisque $Y_{(1)} \neq Y_{(n)}$, chaque fois que m est fixé, on peut choisir ε_o de manière à ce que :

$\forall \varepsilon$, $0<\varepsilon<\varepsilon_o$, $\exists u_1$ et u_2 , $u_1 \neq u_2$ tels que :

$$\begin{cases} \mathrm{Card}(I^{\varepsilon}_{u_1,m} \cap \{y_1,\ldots,y_n\}) \geq 1 \\ \mathrm{Card}(I^{\varepsilon}_{u_2,m} \cap \{y_1,\ldots,y_n\}) \geq 1 \end{cases}$$

On peut également choisir ε_o de manière à ce que $K^{\varepsilon}_{u,m}(\underline{Y})$ soit constant sur $]o,\varepsilon_o]$.

En conclusion, le problème de minoration revient à résoudre :

$$\begin{cases} k \geq 2 \\ \lambda_i \in \mathbb{N}^* \\ \sum_{1}^{k} \lambda_i = n \end{cases}$$

On voit immédiatement qu'un tel problème se ramène au problème :

$$\begin{cases} \text{Max Var } \lambda \\ E(\lambda) = 1 \end{cases}$$

où les moments considérés sont ceux des observations empiriques λ_i , i = 1, ... ,k avec $\lambda_i \in \mathbb{N}^*$ et $k \geq 2$.

Ce problème possède une solution comme on peut le voir dans la partie consacrée à l'analyse de concordance :
en supposant $\lambda_i \leq \lambda_j$ lorsque i < j, cette solution (unique) est donnée par :
$k = 2$; $\lambda_1 = 1$; $\lambda_2 = n-1$.
D'où, au minimum, $n - \frac{1}{n}\sum_{1}^{k} \lambda_i^2 = \frac{2(n-1)}{n}$. ■

3.2 GENERALISATION ET INVARIANCE

On a vu que C(Y) était défini comme une limite, pour ε tendant vers 0, d'une fonction d'indicatrices sur des intervalles du type $I^{\varepsilon}_{u,m}$. $I^{\varepsilon}_{u,m}$ est lui-même extrait d'une partition de $I^{\varepsilon}_{y} = [Y_{(1)} , Y_{(n)} + \varepsilon[$.

3.2.1 Généralisation

On peut généraliser le coefficient C(Y) à un coefficient D(Y) construit à partir d'un intervalle $J_y = [a,b[$, $(a,b \in \mathbb{R} , a \neq b)$, semi-ouvert à droite, tel que :

$$[a,b[\supset [Y_{(1)},Y_{(n)}]$$

J_y est partitionné, comme précédemment, à l'aide d'intervalles notés $J_{u,m}$:

$$J_{u,m} = \left[a + \frac{u-1}{n^m}(b-a) , a + \frac{u}{n^m}(b-a) \right[$$

D(Y) est calculé en fonction de C(Y) en remplaçant I^{ε} par J dans toutes les expressions et en supprimant, par conséquent, les passages à la limite sur ε.

Proposition 1

Si $\mathbb{P}(a,b)$ n'est pas vérifiée alors $D(\underline{Y}) \geq \frac{n+2}{n^2}$

En effet,

on sait que $\forall m \geq 2$,

$$\frac{n_m}{n^m} \geq \frac{2(n-1)}{n^{m+1}}$$

Donc

$$\sum_2^\infty \frac{n_m}{n^m} \geq \frac{2(n-1)}{n^3} \sum_0^\infty \frac{1}{n^m} = \frac{2}{n^2}$$

De plus, comme $\mathbb{P}(a,b)$ n'est pas vérifiée, nécessairement $n_1 \geq 1$.
D'où

$$D(\underline{Y}) \geq \frac{1}{n} + \frac{2}{n^2} = \frac{n+2}{n^2} \quad \blacksquare$$

Proposition 2

$$D(\underline{Y}) \geq 1 - \frac{1}{n} + \frac{2}{n^2} - \frac{y_{(n)} - y_{(1)}}{b-a}$$

En effet,

$[a, y_{(1)}[$ et $]y_{(n)}, b[$ ne contiennent pas d'élément de $(y_1, \cdots, y_n)$ et la somme de leur diamètre est $b-a-y_{(n)}+y_{(1)}$.

On va minorer le nombre d'intervalles de la suite $(J_{u,1})_{u=1,\ldots,n}$ ne contenant aucun élément de (y_i).

Pour cela, soit k le plus grand entier tel que :

$$b-a-y_{(n)}+y_{(1)} \geq k \ \frac{b-a}{n}$$

On a :

$$n - n\frac{y_{(n)} - y_{(1)}}{b-a} \geq k$$

et, par définition de k, on a l'inégalité stricte :

$$k > n-1 - n\frac{y_{(n)} - y_{(1)}}{b-a}$$

Cette borne pour k est un minorant pour n_1.

Et comme $\frac{2}{n^2}$ est un minorant pour $\sum_2^\infty \frac{n_m}{n^m}$, on a bien :

$$D(\underline{Y}) \geq 1 - \frac{1}{n} + \frac{2}{n^2} - \frac{y_{(n)} - y_{(1)}}{b-a} \quad \blacksquare$$

Remarque 1
==========

La borne de la minoration ci-dessus n'est pas nécessairement atteinte. Cependant, pour l'étalement des valeurs de D sur [0,1] on pourra préférer considérer non pas $D(\underline{Y})$ mais $\frac{D(\underline{Y}) - \mu}{1 - \mu}$ où :

$$\mu = 1 - \frac{1}{n} + \frac{2}{n^2} - \frac{y_{(n)} - y_{(1)}}{b-a}$$

Dans ce cas, 1 est une valeur atteinte lorsque la suite (y_i) est constante. Et 0 n'est pas nécessairement atteinte. ■

Remarque 2
==========

On peut envisager les deux cas suivants :

. $\forall i,j \ , \ y_i = y_j$
Dans ce cas, $C(\underline{Y}) = D(\underline{Y}) = 1$

. $\exists i,j \ , \ y_i \neq y_j$
Alors on peut trouver γ et γ' réels tels que :

$$a = y_{(1)} - \gamma \ (y_{(n)} - y_{(1)})$$

$$b = y_{(n)} + \gamma'(y_{(n)} - y_{(1)})$$

($\gamma \geq 0$ et $\gamma' > 0$). ■

3.2.2 Classe d'invariance

Pour savoir à quelle classe d'invariance appartiennent C ou D, on a besoin d'abord de définir une nouvelle manière de calculer $C(\underline{Y})$.

On considère pour ε strictement positif,

$$\tilde{I}^{\varepsilon}_{y} =]y_{(1)} - \varepsilon \ , \ y_{(n)}]$$

et :

$$\tilde{I}^{\varepsilon}_{u,m} =]y_{(1)} - \varepsilon + \frac{u-1}{n^m}(y_{(n)} - y_{(1)} + \varepsilon), y_{(1)} - \varepsilon + \frac{u}{n^m}(y_{(n)} - y_{(1)} + \varepsilon)]$$

$1 \leq u \leq n^m$

En remplaçant I par $\tilde{I}$ dans les définitions de $K^{\varepsilon}_{u,m}(\underline{Y})$ et $N^{\varepsilon}_{u,m}(\underline{Y})$, on obtient $\tilde{K}^{\varepsilon}_{u,m}(\underline{Y})$ et $\tilde{N}^{\varepsilon}_{u,m}(\underline{Y})$.

Les nouvelles fonctions :

$$\tilde{\varphi}_1 : \varepsilon \longmapsto \tilde{N}^{\varepsilon}_{u,m}(\underline{Y})$$

$$\tilde{\varphi}_2 : \varepsilon \longmapsto \tilde{K}^{\varepsilon}_{u,m}(\underline{Y})$$

$$\tilde{\varphi}_3 : \varepsilon \longmapsto \mathrm{Card}(\tilde{I}^{\varepsilon}_{u,m} \cap (y_1, \ldots, y_n))$$

sont telles que les lemmes 1 et 2 sont encore vrais en remplaçant φ_i par $\tilde{\varphi}_i$, i = 1,2,3.

Lemme 4

$$C(\underline{Y}) = \sum_{0}^{\infty} \frac{1}{n^{m+1}} \lim_{\varepsilon \to 0} \sum_{u=1}^{n^m} \frac{\tilde{K}^{\varepsilon}_{u,m}(\underline{Y})}{n} \ \tilde{N}^{\varepsilon}_{u,m}(\underline{Y})$$

En effet, il est clair que $\forall m \geq 0$,

$$n_{m+1} = \lim_{\varepsilon \to 0} \sum_{u=1}^{n^m} \frac{\tilde{K}^{\varepsilon}_{u,m}(\underline{Y})}{n} \ \tilde{N}^{\varepsilon}_{u,m}(\underline{Y}) \quad . \quad ■$$

Remarque
========

Les fonctions $\tilde{\varphi}_i$, i = 1,2,3, sont continues à droite. ■

Proposition 1

Le coefficent C appartient à la classe 3 (cf. les classes d'indices p. 46).

En effet,
si λ_1 et λ_2 sont 2 réels tels que $\lambda_1 \neq 0$, on a :

1) si $\lambda_1 > 0$ et si $y_i \in (y_1, \ldots, y_n)$, $\forall m \geq 0$,

$$y_i \in \left[y_{(1)} + \frac{u-1}{n^m} (y_{(n)} - y_{(1)} + \varepsilon) \ , \ y_{(1)} + \frac{u}{n^m} (y_{(n)} - y_{(1)} + \varepsilon) \right[$$

est équivalent à

$$\lambda_1 y_i \in \left[\lambda_1 y_{(1)} + \frac{u-1}{n^m} (\lambda_1 y_{(n)} - \lambda_1 y_{(1)} + \lambda_1 \varepsilon) \ , \right.$$
$$\left. \lambda_1 y_{(1)} + \frac{u}{n^m} (\lambda_1 y_{(n)} - \lambda_1 y_{(1)} + \lambda_1 \varepsilon) \right[$$

D'après le lemme 2, il est clair qu'on peut remplacer, dans la définition de $C(\underline{Y})$, une limite en ε par une limite en $\lambda_1 \varepsilon$.

Ainsi

$$C(\underline{Y}) = C(\underline{Y}\lambda_1)$$

2) Si $\lambda_1 < 0$, on utilise le lemme 4 et on aboutit au même résultat.

En ce qui concerne une translation définie par λ_2, $\forall m \geq 0$, on a :

$$y_i \in \left[y_{(1)} + \frac{u-1}{n^m} (y_{(n)} - y_{(1)} + \varepsilon) \ , \ y_{(1)} + \frac{u}{n^m} (y_{(n)} - y_{(1)} + \varepsilon) \right[$$

est équivalent à

$$y_i + \lambda_2 \in \left[y_{(1)} + \lambda_2 + \frac{u-1}{n^m} (y_{(n)} + \lambda_2 - y_{(1)} - \lambda_2 + \varepsilon) \ , \right.$$
$$\left. y_{(1)} + \lambda_2 + \frac{u}{n^m} (y_{(n)} + \lambda_2 - y_{(1)} - \lambda_2 + \varepsilon) \right[$$

Ainsi, en translatant non pas $\underline{Y}$ mais $\underline{Y}\lambda_1$, on a :

$$C(\underline{Y}) = C(\underline{Y}\lambda_1 + S\lambda_2)$$ ■

Proposition 2

Le coefficient $D_{\gamma\gamma}$, défini par

$$D_{\gamma\gamma}(\underline{Y}) = 1 \quad \text{si } \forall i,j \ , \ y_i = y_j$$

$$D_{\gamma\gamma}(\underline{Y}) = D(\underline{Y}) \quad \text{si} \begin{cases} \exists i \text{ et } j \mid y_i \neq y_j \\ a = y_{(1)} - \gamma (y_{(n)} - y_{(1)}) \\ b = y_{(n)} + \gamma' (y_{(n)} - y_{(1)}) \end{cases}$$

appartient à la classe 3.

En effet,

. Si $\forall i,j \ , \ y_i = y_j$ alors $\forall \lambda_1, \lambda_2 \in \mathbb{R}$, $\forall i,j$,
$\lambda_1 y_i + \lambda_2 = \lambda_1 y_j + \lambda_2$

d'où

$$D_{\gamma\gamma'}(\underline{Y}) = 1 = D_{\gamma\gamma'}(\underline{Y}\lambda_1 + S\lambda_2)$$

. Si $\exists i$ et $j \mid Y_i \neq Y_j$ alors on peut reprendre les arguments utilisés dans la proposition 1. ■

CHAPITRE 4

Le Coefficient de Concentration comme Indice de Projection

4.1 Etude pratique de la suite des points projetés

Le coefficient de concentration est un indice appliqué, par exemple, à des problèmes géographiques d'étude de populations et à des analyses de séries chronologiques.

Dans sa formulation unidimensionnelle telle qu'on l'a introduit, le coefficient C est une fonction de $\underline{Y}$, autrement dit de la suite de points de $\mathbb{R}$, (Y_i).

Revenons au tableau de données X initial et considérons une projection $\alpha'X$ de Ω.

La restriction de C à l'ensemble de ces projections, lorsque α varie, est un indice de projection pour un problème d'exploration de données.

On pourra alors considérer que C est une fonction :

$$\alpha \longmapsto C(\alpha)$$

Par extension immédiate, on s'intéressera à la restriction de D :

$$\alpha \longmapsto D(\alpha)$$

Dans leur forme, aucun de ces indices ne tient explicitement compte de la notion de distance interpoints.

De ce fait, les points éloignés du "coeur" du nuage n'auront pas une influence particulière par rapport aux autres points.

Ces indices, calculés dans une direction de projection donnée, présentent l'intérêt suivant : ils fournissent une sorte de quantification de la répartition des points projetés.

En effet, leurs deux valeurs extrêmes correspondent :
- l'une à une concentration en un seul point des points projetés, pour la valeur maximum égale à 1
- l'autre à une répartition uniforme des points projetés, pour la valeur minimum.

En terme probabiliste, on dirait que l'échantillon des points projetés est généré par une distribution "comprise" entre une distribution de Dirac et la distribution uniforme et que la

concentration est un indicateur pour ce phénomène.

Tout problème d'optimisation utilisant l'un de ces indices comme objectif ne peut être résolu, à l'aide des algorithmes usuels, que si l'objectif en question, possède une expression analytique soit exacte, soit approchée.

C'est ce qui justifiera les développements à venir.

On note que si le tableau de données X comporte plusieurs échantillons d'une même variable, il n'est pas nécessaire de changer d'échelle : par exemple dans $\mathbb{R}^2$, on ne changera pas d'échelle si les deux variables sont des coordonnées sur un repère générateur; la taille d'un échantillon de Z_α étant alors la taille d'une "population" répartie sur un territoire.

Si les variables considérées sont "incommensurables", on pourra faire un changement d'échelle.

Afin de pouvoir comparer entre elles les différentes valeurs de C, lorsque α varie, il serait souhaitable que les différents échantillons $\{z_1, \ldots, z_n\}_\alpha$ aient même amplitude $z_{(n)} - z_{(1)}$.

Ce n'est généralement pas le cas.

Pour pallier cette difficulté, au lieu de prendre C comme indice de projection, on prend D en déterminant les bornes a et b (bornes du premier intervalle du partitionnement) indépendamment de α.

On prend :

$$a = -\max \|\underline{x}_j\|$$
$$b = \max \|\underline{x}_j\| + \delta \qquad (\delta > 0)$$

L'intérêt de δ apparaît dans le fait que,

$\forall \alpha$, $\forall z_i \in (z_1, \ldots, z_n)_\alpha$,

$$z_i \in [\, -\max \|\underline{x}_j\| \,,\, \max \|\underline{x}_j\| + \delta \,[$$

En réalité, il s'avérera plus pratique de modifier a et b, et les z_i , par affinité de manière à calculer D à partir d'un partitionnement de l'intervalle [0,n[.

Le calcul de $D(\underline{Z}_\alpha)$ sur [a,b[peut se ramener au calcul de $D(\underline{Z}_\alpha \frac{n}{b-a} - S\frac{na}{b-a})$ sur $J_y = [0,n[$ en utilisant le résultat établi à la proposition 2, page 65 .

Pratiquement

On suggère de prendre, pour la simplicité des calculs :

$$\delta = \frac{2 \max \|\underline{x}_j\|}{n}$$

D'où : $b-a = 2\frac{n+1}{n} \max \|\underline{x}_j\|$.

Et :

$$\underline{Z}_\alpha \frac{n}{b-a} - S\frac{na}{b-a} = \underline{Z}_\alpha \frac{n^2}{2(n+1)\max \|\underline{x}_j\|} + \frac{Sn^2}{2(n+1)}$$

On pose : $Y_i = z_i \frac{n^2}{2(n+1)\max \|\underline{x}_j\|} + \frac{n^2}{2(n+1)}$

On calculera

$$D(\underline{Y}) = \sum_1^\infty \frac{n_m}{n^m}$$

où $\underline{Y}$ est défini ci-dessus par ses composantes.

Ensuite, on pourra s'intéresser, en fonction de la remarque faite page 61 , au coefficient :

$$\frac{D(\underline{Y}) - \mu}{1 - \mu}$$

où $\mu = 1 - \frac{1}{n} + \frac{2}{n^2} - (z_{(n)} - z_{(1)}) \frac{n}{2(n+1)\max \|\underline{x}_j\|}$

(en effet, $\forall i$, $y_i \in [0,n[$).

4.2 APPROXIMATION DE $\frac{n_m}{n^m}$ ET DU COEFFICIENT DE CONCENTRATION

On rappelle que, pour $1 \leq u \leq n^m$ et $m \geq 0$,

- $J_{u,m}$ est une maille-m d'un partionnement de $[0,n[$ en n^m intervalles contigus, de même diamètre, semi-ouverts à droite :

$$J_{u,m} = \left[\frac{u-1}{n^{m-1}} , \frac{u}{n^{m-1}} \right[$$

- $K_{u,m}(\underline{Y}) = \sum_{i=1}^{n} 1_{J_{u,m}}(y_i)$

- $N_{u,m}(\underline{Y}) = n - \sum_{v=(u-1)n+1}^{un} 1_{\mathbb{N}^*} \left(\sum_{i=1}^{n} 1_{J_{v,m+1}}(y_i) \right)$

- $n_{m+1} = \sum_{u=1}^{n^m} \frac{K_{u,m}(\underline{Y})}{n} N_{u,m}(\underline{Y})$

Pour trouver $\frac{n_{m+1}}{n^{m+1}}$, il s'agit de trouver $K_{u,m}(\underline{Y})$ et $N_{u,m}(\underline{Y})$ pour $m \geq 0$.

La méthode utilisée pour trouver $\frac{n_{m+1}}{n^{m+1}}$ conduira à des approximations de $K_{u,m}(\underline{Y})$ et $N_{u,m}(\underline{Y})$ pour $m \geq 0$ (à cette exception près que pour $K_{1,0}(\underline{Y})$ on prend sa valeur exacte qui est n).

Pour trouver de façon générale $K_{u,m}(\underline{Y})$ et $N_{u,m}(\underline{Y})$, on a recours à une transformation affine à la fois de $\underline{Y}$ et des bornes de $J_{u,m}$ de façon à se ramener de $J_{u,m}$ à $J_y = [0,n[$.

Ainsi, on pose

$$\underline{W} = (w_1, \ldots, w_n)' = \underline{Y}n^m - Sn(u-1)$$

On a:

$$K_{u,m}(\underline{Y}) = \sum_{i=1}^{n} 1_{[0,n[}(w_i) = \sum_{i=1}^{n} 1_{J_{1,0}}(w_i)$$

$$N_{u,m}(\underline{Y}) = n - \sum_{v=1}^{n} 1_{\mathbb{N}^*} \left(\sum_{i=1}^{n} 1_{J_{v,1}}(w_i) \right)$$

où

$$\underline{W} = \underline{Y}n^m - Sn(u-1) = Z_\alpha \frac{n^{m+2}}{2(n+1)\max\|\underline{x}_j\|} + S\left[\frac{n^{m+2}}{2(n+1)} - n(u-1) \right]$$

N.B.

Désormais, on pourra donc considérer que $K_{u,m}$ et $N_{u,m}$ sont des fonctions de $\underline{W}$.

4.2.1 Approximation de $K_{u,m}(\underline{W})$ et de $N_{u,m}(\underline{W})$

On veut connaître le nombre $K_{u,m}(\underline{W})$ de w_i appartenant à $[0,n[$.

Ceci revient à savoir combien de $\frac{2w_i}{n} - 1$ appartiennent à $[-1,1[$.

Pour cela considérons la suite de fonctions réelles $(g_N)_{\mathbb{N}}$ définie par :

$$g_N(t) = \frac{(t-1)^2}{(t-1)^2+(t+1)^2t^{2N}}$$

(On remarque que $\forall t \in \mathbb{R}$, $(t-1)^2+(t+1)^2t^{2N} \neq 0$)

Propriétés

. $\forall t$, $|t| > 1$, $(g_N(t))_{\mathbb{N}}$ converge vers 0.

. $\forall N$, $g_N(-1) = 1$

. $\forall t$, $|t| < 1$, $(g_N(t))_{\mathbb{N}}$ converge vers 1.

. $\forall N$, $g_N(1) = 0$ ■

Les propriétés concernant g_N permettent d'affirmer que

$$K_{u,m}(\underline{W}) = \lim_{N\to\infty} \sum_{i=1}^{n} g_N\left(2\frac{w_i}{n} - 1\right)$$

c'est-à-dire :

$$K_{u,m}(\underline{W}) = \lim_{N\to\infty} \sum_{i=1}^{n} \frac{1}{1+\left(\frac{w_i}{w_i-n}\right)^2\left(\frac{2w_i-n}{n}\right)^{2N}}$$

De cette égalité on tire, pour "N assez grand", une approximation de $K_{u,m}(\underline{W})$:

$$K_{u,m}(\underline{W}) \cong \sum_{i=1}^{n} \frac{1}{1+\left(\frac{w_i}{w_i-n}\right)^2\left(\frac{2w_i-n}{n}\right)^{2N}}$$

Une approximation de $N_{u,m}(\underline{W})$

On veut connaître le nombre de sous-intervalles de $[0,n[$ du type $[k , k+1[$, $k \in \{0, \ldots , n-1\}$ ne contenant aucun point de (w_i).

On a évidemment l'inégalité :

$$K_{u,m}(\underline{W}) \leq n$$

et il y a au moins $n - K_{u,m}(\underline{W})$ intervalles de la suite $([k , k+1[)_{k=0, \ldots ,n-1}$ ne contenant aucun points de (w_i).

D'autre part, si $[k , k+1[$ contient ξ_k points de (w_i), il y a, si $\xi_k \neq 0$, $\xi_k - 1$ intervalles supplémentaires à ajouter aux $n - K_{u,m}(\underline{W})$ précédents pour obtenir $N_{u,m}(\underline{W})$.

Autrement dit :

Si F est l'ensemble des k tels que $[k , k+1[$ contienne au moins un point de (w_i) et si $\bar{F}$ est l'ensemble complémentaire dans $\{0, \ldots , n-1\}$,

$$N_{u,m}(\underline{W}) = n - K_{u,m}(\underline{W}) + \sum_{F} (\xi_k - 1)$$

Et comme $\xi_k = 0$ lorsque $k \in \bar{F}$,

$$N_{u,m}(\underline{W}) = n - K_{u,m}(\underline{W}) + \sum_{F} (\xi_k - 1) + \sum_{\bar{F}} \xi_k$$

Ce qu'on peut écrire :

$$N_{u,m}(\underline{W}) = n - K_{u,m}(\underline{W}) + \sum_{k=0}^{n-1} (\xi_k - 1_{\mathbb{N}^*}(\xi_k))$$

Il reste à calculer ξ_k et $1_{\mathbb{N}^*}(\xi_k)$.

1) ξ_k s'obtient par une approximation du type de celle utilisée pour $K_{u,m}(\underline{W})$.

On veut connaître le nombre de w_i appartenant à $[k , k+1[$.

Ceci revient à savoir combien de $2w_i-2k-1$ appartiennent à $[-1,1[$.

D'où : $$\xi_k = \lim_{N\to\infty} \sum_{i=1}^{n} g_N(2w_i-2k-1)$$

Pour "N assez grand" :

$$\xi_k \cong \sum_{i=1}^{n} \frac{(w_i-k-1)^2}{(w_i-k-1)^2+(w_i-k)^2(2w_i-2k-1)^{2N}}$$

2) Il est facile de vérifier que, puisque $0 \le \xi_k \le n$,

$$1_{\mathbb{N}^*}(\xi_k) = 1 - \frac{1}{n!} \prod_{h=0}^{n-1} (n-h-\xi_k)$$

En définitive, pour "N assez grand",

$$N_{u,m}(\underline{W}) \cong -K_{u,m}(\underline{W}) + \sum_{k=0}^{n-1} \sum_{i=1}^{n} \frac{(w_i-k-1)^2}{(w_i-k-1)^2+(w_i-k)^2(2w_i-2k-1)^{2N}}$$

$$+ \frac{1}{n!} \sum_{k=0}^{n-1} \prod_{h=0}^{n-1} \left(n-h- \sum_{i=1}^{n} \frac{(w_i-k-1)^2}{(w_i-k-1)^2+(w_i-k)^2(2w_i-2k-1)^{2N}} \right)$$

4.2.2 Approximation de D(α)

On sait (cf. page 73) que w_i est fonction de u et de m ($1 \le u \le n^m$) ainsi que de α via $\underline{Z}_\alpha = \alpha'X$.

On pose

$$G_N(\alpha,u,m,i) = \frac{1}{1+\left(\frac{w_i}{w_i-n}\right)^2\left(\frac{2w_i-n}{n}\right)^{2N}}$$

$$H_N(\alpha,u,m,i,k) = \frac{(w_i-k-1)^2}{(w_i-k-1)^2+(w_i-k)^2(2w_i-2k-1)^{2N}}$$

Ainsi : $$K_{u,m}(\underline{W}) \cong \sum_{i=1}^{n} G_N(\alpha,u,m,i)$$

$$N_{u,m}(\underline{W}) + K_{u,m}(\underline{W}) \cong \sum_{k=0}^{n-1} \sum_{i=1}^{n} H_N(\alpha,u,m,i,k) + \frac{1}{n!} \sum_{k=0}^{n-1} \prod_{h=0}^{n-1} \left[n-h- \sum_{i=1}^{n} H_N(\alpha,u,m,i,k) \right]$$

$$n_{m+1} = \lim_{N\to\infty} \sum_{u=1}^{n^m} \frac{N_{u,m}(\underline{W})K_{u,m}(\underline{W})}{n}$$

$$= \lim_{N\to\infty} \sum_{u=1}^{n^m} -\frac{1}{n}K^2_{u,m}(\underline{W}) + \frac{1}{n}K_{u,m}(\underline{W})(K_{u,m}(\underline{W})+N_{u,m}(\underline{W}))$$

$$= \lim_{N\to\infty} \frac{1}{n}\sum_{u=1}^{n^m}\left\{-\left(\sum_{i=1}^{n}G_N\right)^2+\left(\sum_{i=1}^{n}G_N\right)\left(\sum_{k=0}^{n-1}\sum_{i=1}^{n}H_N+\frac{1}{n!}\sum_{k=0}^{n-1}\prod_{h=0}^{n-1}\left[n-h-\sum_{i=1}^{n}H_N\right]\right)\right\}$$

(en simplifiant l'écriture).

D'autre part, d'après la proposition de la page 56, si $M \geq 0$,

$$\sum_{M+1}^{\infty}\frac{n_{m+1}}{n^{m+1}} = \sum_{M+2}^{\infty}\frac{n_m}{n^m} \geq \sum_{M+2}^{\infty}\frac{2(n-1)}{n^{m+1}}$$

Ainsi $\sum_{M+1}^{\infty}\frac{n_{m+1}}{n^{m+1}}$ est minoré par $\frac{2}{n^{M+2}}$ et majoré, d'après le lemme 3 (p. 52) par $\frac{1}{n^{M+1}}$.

D'où l'on tire une estimation moyenne pour $\sum_{M+1}^{\infty}\frac{n_{m+1}}{n^{m+1}}$:

$$\frac{1}{2}\left(\frac{2}{n^{M+2}}+\frac{1}{n^{M+1}}\right) = \frac{n+2}{2n^{M+2}}$$

Donc, "pour N assez grand",

$$D(\alpha) \cong \sum_{m=0}^{M}\frac{1}{n^{m+2}}\sum_{u=1}^{n^m}\left\{-\left(\sum_{i=1}^{n}G_N\right)^2+\left(\sum_{i=1}^{n}G_N\right)\left(\sum_{k=0}^{n-1}\sum_{i=1}^{n}H_N+\frac{1}{n!}\sum_{k=0}^{n-1}\prod_{h=0}^{n-1}\left[n-h-\sum_{i=1}^{n}H_N\right]\right)\right\}+\frac{n+2}{2n^{M+2}}$$

4.3 ETUDE DE LA STRUCTURE DE Ω A L'AIDE DU COEFFICIENT DE CONCENTRATION

On désire déceler les structures de Ω en optimisant D par rapport à α.

La démarche employée est de type B (cf. page 14).

On cherche le "meilleur" hyperplan de projection de $\mathbb{R}^p$ susceptible d'inclure une bonne représentation de Ω.

Une manière de procéder consiste, comme en analyse factorielle classique, à rechercher les uns après les autres, les axes "de plus grand allongement" du nuage (C., Raffestin et C., Tricot, (1975), parlent d'indice d'allongement).

En analyse factorielle classique, on maximise sur chacun des axes la variance des composantes principales.

Ici, les considérations dégagées auparavant, concernant la répartition des points projetés (cf. page 68), amènent à minimiser D sur chacun des axes.

NOTE : Si $D(\alpha)$ est calculé sur un tableau X, on prendra plutôt la notation non simplifiée : $D(\alpha'X)$.

1ère étape :

On recherche l'axe α_1 tel que les projections des points entre 0 et n sur cet axe aient "une occupation maximum", ou si l'on veut, soient réparties "le plus uniformément possible" (au sens qu'on a précisé lorsqu'on a étudié l'indice D).

α_1 est la solution du problème :

$$\begin{cases} \min D(\alpha'X) \\ \alpha'\alpha = 1 \end{cases}$$

2ème étape :

On projette Ω dans le sous-espace orthogonal à α_1 à l'aide de la matrice idempotente $I - \alpha_1\alpha_1'$.

On forme ainsi un nouveau tableau de données noté X_1 :

$$X_1 = (I - \alpha_1\alpha_1')X$$

(L'idée exploitée est celle de recherches de structures par élimination des structures déjà obtenues : Friedman, 1987).

3ème étape :

On recherche α_2, solution de :

$$\begin{cases} \min D(\alpha'X_1) \\ \alpha'\alpha = 1 \\ \alpha'\alpha_1 = 0 \end{cases}$$

Remarque 1
==========

Il faut noter que nécessairement on doit avoir :

$$D(\alpha_1'X_1) = 1$$

puisque l'espace orthogonal à α_1 est projeté en 0 sur α_1.

Cependant il ne faut pas perdre de vue qu'un programme d'objectif D est résolu à l'aide d'une approximation de D : il se peut donc, en particulier, que $D(\alpha_1'X_1)$ ne soit pas exactement égal à 1. ■

Remarque 2
==========

Si n est "assez grand", on peut s'attendre, en résolvant simplement :

$$\begin{cases} \min D(\alpha' X_1) \\ \alpha'\alpha = 1 \end{cases}$$

à obtenir une solution "proche" de l'espace orthogonal à α_1 et donc "proche" de α_2.

Compte tenu du fait que l'objectif est toujours approximé, il pourra donc apparaître plus simple et aussi efficace de trouver une direction de concentration minimum orthogonale à α_1 en résolvant le programme ci-dessus (on pourra éventuellement faire une correction sur la direction trouvée pour obtenir une direction orthogonale à α_1). ■

Remarque 3
==========

Si X peut être considéré comme un échantillon d'un vecteur normal à p dimensions, il pourra s'avérer intéressant de traduire des phénomènes d'orthogonalité en terme d'indépendance en probabilité et ceci, grâce à l'orthogonalité des directions de projection. ■

4ème étape :

On forme le nouveau tableau :

$$X_2 = (I-\alpha_1\alpha_1'-\alpha_2\alpha_2')X$$

où la matrice de la transformation de X est idempotente et projette les points de Ω sur un espace orthogonal au plan engendré par α_1 et α_2.

5ème étape :

En fonction de la remarque précédente, on résoud ou bien :

$$\begin{cases} \min D(\alpha' X_2) \\ \alpha'\alpha = 1 \\ \alpha'\alpha_1 = 0 \\ \alpha'\alpha_2 = 0 \end{cases}$$

ou bien simplement :

$$\begin{cases} \min D(\alpha' X_2) \\ \alpha'\alpha = 1 \end{cases}$$

etc... .

On calcule ainsi successivement $\alpha_1, \alpha_2, \ldots, \alpha_{n-1}$.

Remarque 4
==========

Au terme de ces calculs, on pourra comparer la direction orthogonale aux α_i, i = 1, ..., n-1, à la direction obtenue en résolvant le programme :

$$\begin{cases} \max D(\alpha'X) \\ \alpha'\alpha = 1 \end{cases} \quad \blacksquare$$

4.3.1 Projection en dimension 2

La remarque précédente nous amène, en dimension 2 (α : 2x1), à faire la conjecture que, au sens par exemple de la norme euclidienne $\| \ \|^2$, les couples de solutions des trois problèmes suivants sont proches :

1) $\begin{cases} \min D(\alpha'X) \\ \alpha'\alpha = 1 \end{cases}$ et $\begin{cases} \max D(\alpha'X) \\ \alpha'\alpha = 1 \end{cases}$

2) $\begin{cases} \min [D(\alpha'X) - D(\beta'X)] \\ \alpha'\alpha = 1 \\ \beta'\beta = 1 \end{cases}$

3) $\begin{cases} \min [D(\alpha'X) - D(\beta'X)] \\ \alpha'\alpha = 1 \\ \beta'\beta = 1 \\ \alpha'\beta = 0 \end{cases}$

4.3.2 Forme du nuage dans le plan

Rechercher la forme du nuage dans un plan signifie procéder en deux étapes :

La première étape consiste à trouver le plan de projection en dimension 2 rendant compte le mieux possible de la structure homogène du nuage Ω.

Pour cela on adopte les points 1),2) et 3) de la construction présentée précédemment ce qui donne deux directions α_1 et α_2 générant un plan.

C'est à ce plan que l'on se référera dans la deuxième étape.

La deuxième étape consiste à trouver la ligne, dans le plan précédent, la plus "proche" du nuage projeté dans ce plan.

Cette proximité va être définie à l'aide des deux principes de construction suivants :

Précisons d'abord qu'on s'intéresse à une ligne polygonale et notons X^o le tableau de données 2xn dont on dispose et X^1 un tableau extrait de X^o par suppression de colonnes.

Premier principe de construction :

- la ligne doit être une ligne médiane pour le nuage Ω_o correspondant à X^o.

Deuxième principe de construction :

- Chaque segment, obtenu au vu d'une partie Ω_1 des points de Ω_o (correspondant à un sous-tableau X^1) doit avoir une direction β_1 ($\|\beta_1\|=1$) qui minimise $D(\beta' X^1)$. (On envisagera autant de parties de Ω_o que de segments de la ligne polygonale).

1ère approche : la ligne est constituée d'un seul segment.

On résoud :

$$\begin{cases} \min [D(\beta' X^o) - D(\gamma' X^o)] \\ \beta'\beta = 1 \\ \gamma'\gamma = 1 \\ \beta'\gamma = 0 \end{cases}$$

Les solutions sont notées β_o et γ_o.

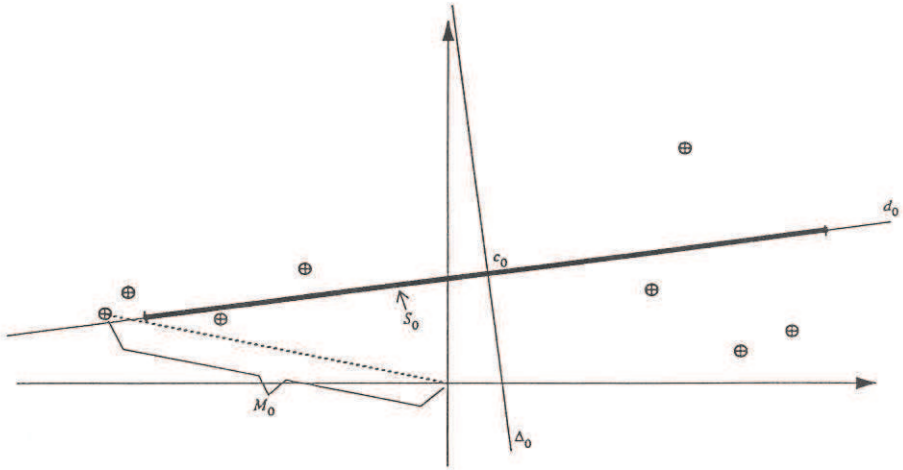
On considère les deux droites d_o et Δ_o , de direction respectivement β_o et γ_o , qui soient toutes les deux médianes du nuage Ω_o.

Elles se coupent en un point c_o.

Soit M_o la plus grande des normes des colonnes de X^o.

On prend comme ligne, le segment de droite S_o d'extrêmités :

$$M_o\beta_o + c_o \qquad \text{et} \qquad -M_o\beta_o + c_o$$



2ème approche : la ligne est constituée de deux segments.

Δ_o sépare Ω_o en deux nuages Ω_1 et Ω_2 (si un point de Ω_o est sur Δ_o , on admettra qu'il appartient à la fois à Ω_1 et Ω_2).

Ω_1 correspond à un tableau X^1.
Ω_2 correspond à un tableau X^2.

On résoud pour i = 1,2 ,

$$\begin{cases} \min [D(\beta' X^i) - D(\gamma' X^i)] \\ \beta'\beta = 1 \\ \gamma'\gamma = 1 \\ \beta'\gamma = 0 \end{cases}$$

Pour i = 1 on obtient les solutions β_1 et γ_1.
Pour i = 2 on obtient les solutions β_2 et γ_2.

On trouve c_1 , le point d'intersection des droites d_1 et Δ_1 de direction β_1 et γ_1 tel que ces droites soient médianes pour Ω_1.

On trouve, de même avec Ω_2 , le point c_2 (les droites sont d_2 et Δ_2).

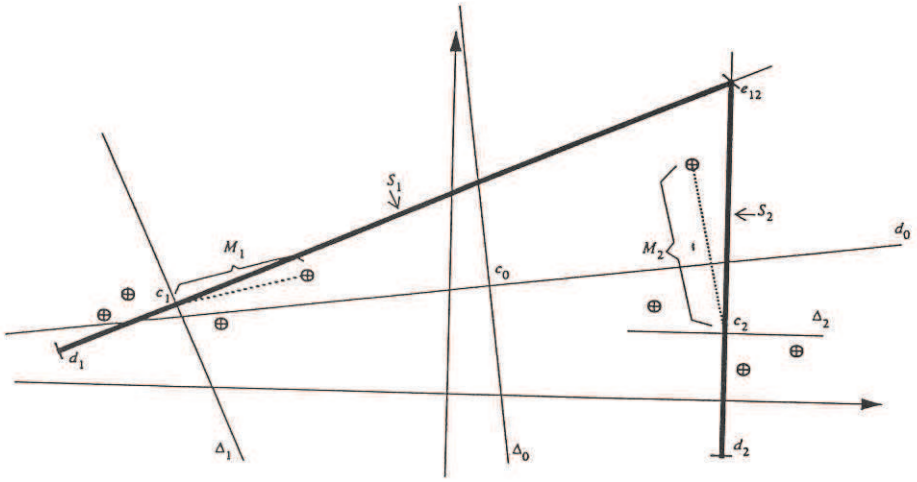
Notons e_{12}, le point d'intersection des droites (d_1,c_1) et (d_2,c_2).

Notons M_1 (resp. M_2) la plus grande des normes des colonnes de X^1-c_1 (resp. X^2-c_2).

On prend le segment S_1 d'extrêmités $M_1\beta_1+ c_1$ et e_{12} si c_1 lui appartient, et $-M_1\beta_1+ c_1$ et e_{12} sinon.

On fait de même avec l'indexation 2 et l'on obtient S_2.

On prend, comme ligne polygonale, la réunion $S_1 \cup S_2$ $(S_1 \cap S_2 = e_{12})$.



Les approches suivantes s'obtiennent en décomposant encore Ω_1 et Ω_2 : la $r^{ième}$ approche prend en compte 2^r nuages dont la réunion forme Ω_o.

Ceci donne lieu a une ligne polygonale composée de 2^r segments de droite.

Le critère d'arrêt de la méthode sera, par exemple, un seuil sur le cardinal des différents nuages de la décomposition de Ω_o.

Une telle ligne polygonale peut mettre en évidence certains regroupements des points de Ω_o en classes de points.

* * *

DEUXIEME PARTIE

ANALYSE DE CONCORDANCE

ET APPLICATION DE LA METHODE DES RESEAUX

CHAPITRE 1

Analyse de Concordance : Cadre et Exemples

1.1 CADRE D'UNE ANALYSE DE CONCORDANCE

Il est utile d'indiquer les termes permettant de préciser le cadre d'une telle analyse :

Une analyse de concordance autrement dénommée par Beckett et Schucany (1975) ANACONDA (par analogie avec une ANOVA, du fait que les méthodes employées dans l'un et l'autre domaine sont souvent proches) met en rapport trois types de grandeurs appelées : Sujet, évaluation et observateur (voir, par exemple, Schouten, 1980).

Parfois, dans la littérature, un sujet sera appelé un objet, une évaluation, un choix et un observateur, un juge.

On suppose qu'il y a d observateurs, $d \geq 2$ (indice j), et n

sujets, n≥1 (indice i). Il faut ensuite définir une variable qui pourra être quantitative ou qualitative (nominale ou ordinale).
Les observateurs évaluent chacun des sujets sur l'échelle des valeurs possibles prises par la variable :

Chaque observateur est donc amené à faire n évaluations, une évaluation par sujet, et une analyse de concordance consiste en une analyse du degré d'accord entre les d observateurs, c'est-à-dire, du degré de similitude entre d vecteurs nx1 d'évaluations (la $i^{ème}$ composante de chaque vecteur correspond à une évaluation du sujet i).

On peut dresser le tableau X, nxd, comportant en colonne ces d vecteurs, et noter x_{ij} l'élément générique de X.

On étudiera, par la suite, le cas où la variable est qualitative nominale. Cependant, il existe un lien entre les différentes méthodes connues utilisées lorsque la variable est quantitative et lorsqu'elle est qualitative.

C'est pourquoi on va indiquer maintenant quelques-unes de ces méthodes dans les différents cas.

1.2 EXEMPLES DE MESURES EN PRESENCE DES DIFFERENTS TYPES DE DONNEES DISPONIBLES

1.2.1 Les évaluations sont des valeurs prises par une variable quantitative

On suppose que la $i^{ème}$ ligne du tableau X est générée par un vecteur aléatoire :

$$X_i = (X_{i1}, \ldots, X_{id})' \qquad (dx1)$$

Autrement dit, x_{ij} est un échantillon de taille 1 d'une variable aléatoire X_{ij} .

Robinson (1957) et après lui Bartko (1966), se rapportent à des modèles d'analyse de variance (ANOVA) pour mesurer le degré d'accord entre les observateurs.

Ils considèrent, au vu du tableau X, le modèle d'une analyse hiérarchique à 2 générations (dit modèle II) :

$$x_{ij} = \mu + s_i + \varepsilon_{ij}$$

où

- μ est une constante réelle
- s_i , i = 1, ..., n , s'interprète comme un "effet sujet"
- ε_{ij} ,i = 1, ..., n ; j = 1, ..., d , s'interprète comme un "effet résiduel"

Ceci conduit à affiner le modèle dans la mesure où l'on dispose, pour chaque variable X_{ij} d'un r-échantillon :

$$(x_{ij1}, \ldots, x_{ijr})$$

Ce qui revient à dire que l'on dispose d'un tableau de données X_r , nxdxr à r dimensions.

Ainsi, au vu du tableau X_r , on considère le modèle d'une analyse hiérarchique à 3 générations (dit modèle III) :

$$x_{ijk} = \mu + s_i + \delta_j + e_{ijk}$$

où

- μ est une constante réelle
- s_i , i = 1, ..., n , s'interprête comme un "effet sujet"
- δ_j , j = 1, ..., d , s'interprête comme un "effet observateur"
- e_{ijk} , i = 1, ..., n ; j = 1, ..., d ; k = 1, ..., r, s'interprête comme un "effet résiduel"

Dans le cas du modèle II, une mesure du degré d'accord est :

$$R^2 = \frac{(n-1)MS_s}{(n-1)MS_s + n(d-1)MS_e}$$

où

- $MS_s = \frac{d}{n-1} \sum_i (\bar{x}^i - \bar{x})^2$
- $MS_e = \frac{1}{n(d-1)} \sum_{i,j} (x_{ij} - \bar{x}^i)^2$

et

- $\bar{x}^i = \frac{1}{d} \sum_j x_{ij}$

- $\bar{x} = \frac{1}{nd} \sum_{i,j} x_{ij}$

MS_s et MS_e sont, respectivement, des estimations non biaisées d'une variance entre sujets et d'une variance résiduelle, sous de bonnes hypothèses de normalité et d'indépendance sur la matrice de données.

R^2 est exactement le coefficient de corrélation multiple intervenant en analyse de régression linéaire classique.
En effet, on a l'équation d'analyse de variance :

$$\sum_{i,j} (x_{ij} - \bar{x})^2 = \sum_{i,j} (x_{ij} - \bar{x}^i)^2 + d \sum_i (\bar{x}^i - \bar{x})^2$$

Et, par conséquent,

$$R^2 = 1 - \frac{\sum_{i,j} (x_{ij} - \bar{x}^i)^2 / nd}{\sum_{i,j} (x_{ij} - \bar{x})^2 / nd}$$

où $\sum_{i,j} (x_{ij} - \bar{x}^i)^2 /nd$ est une estimation de la variance résiduelle et $\sum_{i,j} (x_{ij} - \bar{x})^2/nd$ est une estimation de la variance totale.

Plus l'effet résiduel qui traduit "le désaccord" entre les observateurs sera petit, plus le degré d'accord sera élevé.
$R^2 = 1$ correspond à l'accord parfait, c'est-à-dire,

$$\forall j,j' , \ \underline{x}_j = \underline{x}_{j'},$$

1.2.2 Les évaluations sont des valeurs prises par une variable qualitative ordinale

On note a_i le $i^{ème}$ sujet (on dit aussi : le sujet i). L'ensemble des sujets $\{a_1, \ldots, a_n\}$ est supposé rangé par chaque observateur : l'ensemble des évaluations possibles est donc l'ensemble $\{1, \ldots, n\}$.

Autrement dit, les observateurs établissent un ordre de préférence entre les sujets et une colonne de X est donc une permutation des éléments de $\{1, \ldots, n\}$ (absence d'ex-aequo).

Comme on lit dans Sibson (1972), la notion de préférence ne tient pas compte des différentes notions de distances entre sujets ou de transformation monotone des sujets; c'est pourquoi l'observateur j établit une correspondance biunivoque entre l'ensemble $\{a_1, \ldots, a_n\}$ des n sujets et l'ensemble discret "le plus simple" $\{1, \ldots, n\}$, à l'aide de la fonction $rang_j$:

L'évaluation de a_i par l'observateur j est :

$$rang_j(a_i) = x_{ij}$$

Les d colonnes de X deviennent ainsi un d-échantillon d'un espace de préférence Ω, fini, contenant n! éléments.

On est amené à mesurer l'accord entre d observateurs au vu d'un d-échantillon tiré de Ω .

L'accord sera parfait si les éléments de ce d-échantillon sont égaux.

La mesure du degré d'accord la plus connue est celle donnée par Kendall (1970). Il s'agit du "coefficient de concordance de Kendall" :

$$W = \frac{12S}{n^2 - 1}$$

où $\quad S = \frac{1}{n} \sum_i (\bar{x}^i - \bar{x})^2.$

Il est clair que l'on a :

$$\bar{x} = \frac{n + 1}{2}$$

et,

$$\frac{1}{n} \sum_i \left(i - \frac{n+1}{2}\right)^2 = \frac{1}{nd} \sum_{i,j} (x_{ij} - \bar{x})^2 = \frac{n^2-1}{12}$$

qui s'interprête comme une variance totale ou comme une variance maximale par référence aux valeurs possibles de S.

Et par conséquent, on a, sur l'espace Ω :

$$W = R^2$$

Pour mémoire, et en lien direct avec le coefficient W, Kendall (1970) construit un autre coefficient en prenant une

moyenne entre $\binom{d}{2}$ coefficients de corrélation de Spearman :

$$\rho_{av} = \frac{2}{d(d-1)} \sum_{\substack{j,j' \\ j<j'}} \left\{ \frac{\sum_i (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})}{n(n^2-1)/12} \right\}$$

où $\bar{x} = \frac{1}{n} \sum_i x_{ij}$.

On montre alors aisément l'égalité : $\rho_{av} = \frac{dW - 1}{d - 1}$.

W est donc linéairement lié à $\binom{d}{2}$ coefficients de corrélation de Spearman.

En commentaire à l'article de Hollander et Sethuraman (1978), Schucany écrit :

> *"Depuis le début des travaux sur le problème de la comparaison de d rangements, une seule mesure adéquate de l'accord se dégage comme étant celle de la correlation des rangs. (...). Les autres mesures se relient directement à cette dernière".*

Malgré cette remarque, Gordon (1979) introduit une autre mesure ; c'est celle que l'on va brièvement exposer maintenant.

On suppose que l'indexation sur les a_i est telle que l'un des observateurs, dit "de référence", effectue le rangement trivial : i est l'évaluation qu'il porte sur a_i .
Soit $(A_1, \ldots, A_g)$ une partition de $(1, \ldots, n)$, $(1 \leq g \leq n)$.

On suppose cette partition telle que :

$\forall i,i'$ appartenant à $(1, \ldots, n)$,
$\forall k,k'$ $k<k'$ appartenant à $(1, \ldots, g)$, on a :
$\forall j \in (1, \ldots, d)$,

$$\left.\begin{array}{l} i \in A_k \\ i' \in A_{k'} \end{array}\right\} \Longrightarrow \text{rang}_j(a_i) < \text{rang}_j(a_{i'})$$

Autrement dit, les observateurs adoptent tous les mêmes évaluations (mais dans un ordre éventuellement distinct) pour les sujets indexés dans une même partie A_k .

Une telle partition $(A_1, \ldots, A_g)$ de $(1, \ldots, n)$ n'est pas nécessairement unique comme le montre l'exemple suivant :

Prenons d = 2 ; n = 5
Considérons les deux ensembles d'évaluations :

1 2 3 4 5 pour l'observateur 1 (observateur de référence)
2 1 3 5 4 pour l'observateur 2

- g = 2 : $A_1 = (1,2,3)$; $A_2 = (4,5)$
Si $i \in A_1$ et $i' \in A_2$, on a bien pour chacun des deux ensembles d'évaluations : rang (a_i) < rang $(a_{i'})$
- g = 3 : $A_1 = (1,2)$; $A_2 = (3)$; $A_3 = (4,5)$
même conclusion.

La mesure de l'accord envisagée est

$$\gamma_n = \max g \qquad (1 \leq g \leq n)$$

où le maximum est obtenu sur un ensemble adéquat de partitions de $(1, \ldots, n)$.

L'accord est parfait, lorsque $\gamma_n = n$.

Ensuite, dans une optique de test, on choisit toujours un observateur comme observateur de référence.

On suppose qu'en supprimant la colonne de X correspondant à cet observateur, le (d-1)-échantillon de l'espace Ω qui résulte du tableau X ainsi tronqué, est un point ω de l'espace Ω^{d-1}.

On fait l'hypothèse H_0 suivante :

Ω^{d-1} possède une distribution uniforme

Autrement dit, $P(\omega) = \frac{1}{(n!)^{d-1}}$.

Une telle hypothèse revient à admettre que les observateurs ne distinguent pas les sujets les uns des autres : chaque observateur effectue donc un rangement des sujets de façon

totalement arbitraire.

On est conduit à considérer le test de l'hypothèse H_0 basé sur une région critique du type $(c, c + 1, \ldots, n)$ à l'aide de la fonction discriminante γ_n , tel que

$$P(\gamma_n \geq c) = \alpha$$

(α, erreur de 1ère espèce).

Interprétation

On interprète le rejet de H_0 comme le rejet d'un certain "degré de désaccord".

Un tel désaccord prend la forme du résultat d'un tirage au hasard, par chaque observateur, d'un ensemble d'évaluations pris parmi les éléments de Ω. ■

1.2.3 Les évaluations sont des valeurs prises par une variable qualitative nominale

On suppose que les évaluations possibles appartiennent à un ensemble noté $E = \{e_1 , \ldots, e_L\}$: on a L catégories.

Puisque la variable est qualitative nominale, cet ensemble est dépourvu de toute structure.

Cependant, par la suite, on considèrera le cas où E est muni d'une structure comme une extension du cas d'absence de structure :

Une structure que l'on envisagera sera une structure de ressemblance (voir, par exemple, Diday, 1982) : ceci revient à se donner sur $\{e_1 , \ldots, e_L\}$ une mesure de ressemblance, c'est-à-dire à disposer d'un tableau de proximité sur l'ensemble d'"individus" que représente $\{e_1 , \ldots, e_L\}$.

D'autre part, le tableau de données X, nxd, n'est pas celui dont on se servira directement à partir de maintenant :

A la manière de Fleiss (1971) (et après lui Schouten, 1980), on dresse le tableau de contingence Y, nxL, déduit du tableau X en reportant à l'intersection de la $i^{ème}$ ligne et de la colonne associée à l'évaluation e , le nombre d'observateurs ayant porté, sur le sujet i, l'évaluation e.

Si n_{ie} est l'élément générique de Y, on a donc :

$\forall i , \sum_{e \in E} n_{ie} = d$

Autrement dit, la somme des éléments d'une ligne de Y est toujours égale au nombre d'observateurs, en faisant l'hypothèse (que l'on pourra abandonner plus tard) que chaque sujet est évalué par le même nombre d'observateurs.

Autre formulation : quelque soit i, la suite $(n_{ie})_e$ des éléments non nuls de la $i^{ième}$ ligne de Y est une partition de l'entier d.

LA MESURE KAPPA

On considère une population infinie d'observateurs et l'on admet ainsi que les d observateurs, retenus dans l'analyse de concordance, constituent un d-échantillon extrait de la population.

Dans cette population, on considère les événements suivants :

(i,e) : "l'évaluation e est portée sur le sujet i"

(e) : "l'évaluation e est portée sur un sujet quelconque"

(i,e,e') : "les évaluations e et e' sont portées sur le sujet i"

(e,e') : "les évaluations e et e' sont portées sur un même sujet quelconque"

sur lesquels on établit une distribution de probabilité.

On note les probabilités de ces différents événements respectivement Π_{ie} ; Π_{e} ; $\Pi_{iee'}$ et $\Pi_{ee'}$.

On a les estimations suivantes :

Une estimation non biaisée de Π_{ie} :

$$\frac{n_{ie}}{d}$$

Une estimation non biaisée de Π_{e} :

$$p_e = \frac{1}{n} \sum_i \frac{n_{ie}}{d} = \frac{n_{.e}}{nd}$$

(en posant $n_{.e} = \sum_i n_{ie}$: élément de la marge du bas de Y).

Une estimation non biaisée de $\Pi_{iee'}$:

Si $e = e'$, $\binom{n_{ie}}{2} / \binom{d}{2}$

Si $e \neq e'$, $n_{ie}\, n_{ie'} / \binom{d}{2}$

Une estimation non biaisée de $\Pi_{ee'}$:

Si $e = e'$, $p_{ee} = \frac{1}{n} \sum_i \binom{n_{ie}}{2} / \binom{d}{2} = \frac{1}{n} \sum_i \frac{n_{ie}\left(n_{ie}-1\right)}{d(d-1)}$

Si $e \neq e'$, $p_{ee'} = \frac{1}{n} \sum_i n_{ie} n_{ie'} / \binom{d}{2} = \frac{1}{n} \sum_i \frac{2 n_{ie} n_{ie'}}{d(d-1)}$

Sous l'hypothèse H définie par Cohen (1960),

> *"Les observateurs sont dans l'incapacité de distinguer les sujets les uns des autres",*

on admet que les évaluations sont portées par les observateurs de manière indépendante, ce qui se traduit par :

$$\Pi_{ee} = \Pi_e^2$$
$$\Pi_{ee'} = 2\Pi_e . \Pi_{e'}, \qquad (e \neq e')$$

et, dans ce cas, Π_{ee} est estimé par p_e^2 et $\Pi_{ee'}$, $(e \neq e')$, par $2p_e p_{e'}$.

On rejoint ici exactement le modèle de désaccord exprimé dans l'approche de Gordon (1979) étudiée plus haut.

Une estimation d'une probabilité d'accord des observateurs 2 par 2 est donc :

$$p^o = \sum_e p_{ee}$$

Cette mesure a été introduite par Cartwright (1956) ; et sous l'hypothèse H, cette estimation devient :

$$p^c = \sum_e p_e^2$$

(appelée, dans la littérature, mesure de l'accord dû à la chance).

La mesure Kappa est définie comme un coefficient de Cartwright "corrigé":

$$\kappa = \frac{p^o - p^c}{1 - p^c}$$

(où le facteur $(1 - p^c)^{-1}$ n'a qu'un effet de normalisation).

Un test de l'hypothèse H devient ainsi celui de l'hypothèse : $\kappa = 0$.

LES MESURES KAPPA PONDEREES

Ces mesures procèdent des techniques précédentes moyennant la prise en compte d'un tableau de proximité Q sur les éléments de $\{ e_1, \ldots, e_L \}$ $(L > 1)$.

Etant données e et e' appartenant à $(e_1, \ldots, e_L)$, l'élément générique $q_{ee'}$ de Q peut prendre différentes formes.
Donnons trois illustrations :

On suppose que $e = e_a$ et $e' = e_b$ $(a, b \in \{1, \ldots, L\})$.

1) $q_{ee'} = |a - b|$ (renoté $q'_{ee'}$)

2) $q_{ee'} = 1 - \frac{q'_{ee'}}{\max q'_{ee'}} = 1 - \frac{|a-b|}{L-1}$

3) $q_{ee'} = (a - b)^2$

On pose :

$$p^{oq} = \sum_{\substack{e,e' \\ e<e'}} \frac{1}{n} \sum_i q_{ee'} n_{ie} n_{ie'} \Big/ \binom{d}{2}$$

$$p^{cq} = \sum_{\substack{e,e' \\ e \neq e'}} q_{ee'} \frac{n_{.e}}{nd} \cdot \frac{n_{.e'}}{nd}$$

Les interprétations de p^{oq} et p^{cq} découlent de celles qui ont été données à p^o et p^c.

On définit alors les mesures suivantes :

$$\kappa_1 = \frac{p^{oq} - p^{cq}}{1 - p^{cq}} \quad \text{(lorsque } \max q_{ee'} = q_{ee}\text{)}$$

qui a la forme de κ ; et une variante :

$$\kappa_2 = \frac{p^{cq} - p^{oq}}{p^{cq}} \quad \text{(lorsque } \min q_{ee'} = q_{ee}\text{)}$$

Schouten (1982) a démontré que κ_2 était formellement égale, dans le cas du troisième exemple ci-dessus, à un coefficient de corrélation empirique (corrélation intraclasse) en considérant $\{ 1, \ldots, L \}$ comme un échantillon généré par une variable aléatoire réelle.

On retrouve donc, dans le cas d'une variable qualitative

comme dans le cas des autres types de variable, une association entre la notion d'accord et la notion de corrélation.

Remarque 1
==========

La mesure κ non pondérée peut être considérée comme une mesure Kappa pondérée du type κ_1 où le tableau de proximité Q sur $\{ e_1, \ldots, e_L \}$ est égal à I. ■

Remarque 2
==========

Il pourra être utile de préciser les conditions dans lesquelles les mesures Kappa sont toujours bien définies, c'est-à-dire, n'admettent pas de dénominateur nul.

On traitera, par exemple, à part le cas où $n_{.e} = nd$ qui implique $1-p^c = 0$. ■

Une interprétation de la proportion d'accord p^o ; lien avec le tau de Kendall moyen

On peut constater que la suite $(n_{ie})_e$, formant une partition de d (lorsque $n_{ie} \neq 0$), est un élément de l'ensemble fini des partitions de d de cardinal δ.

Notons $(p_1 , \ldots, p_\delta)$ cet ensemble de partitions de d.

$\forall i, \exists k \mid (n_{ie})_e = p_k$

On pose :

$$\lambda_k = \text{Card} \{ i, 1 \leq i \leq n \mid (n_{ie})_e = p_k \}$$

On a donc :

$$\sum_{k=1}^{\delta} \lambda_k = n$$

D'autre part, sur le sujet i, la proportion d'accords des observateurs 2 par 2, est égale à

$$\sum_e \binom{n_{ie}}{2} / \binom{d}{2}$$

On note α_k cette proportion (dans la mesure où $(n_{ie})_e = p_k$).

On voit que $p^o = \sum_e \frac{1}{n} \sum_i \binom{n_{ie}}{2} / \binom{d}{2} = \frac{1}{n} \sum_i \sum_e \binom{n_{ie}}{2} / \binom{d}{2}$

est le barycentre des α_k pour les poids λ_k/n.

Rappelons maintenant une définition du tau de Kendall moyen :

On utilise ici une définition qui s'inspire de la définition rapportée par Kendall (1970), lorsqu'il compare entre eux les ordres issus de plusieurs rangements.

Soit la relation d'équivalence entre les observateurs à

propos du sujet i :

R_i : "porte la même évaluation sur i que"

On rappelle que j est utilisé pour indexer les observateurs.

On pose :

$$P_{ijj'} = \begin{cases} 1 & \text{si } jR_ij' \text{ et } j \neq j' \\ 0 & \text{sinon} \end{cases}$$

$$Q_{ijj'} = 1 - P_{ijj'}$$

Sur le sujet i, on prend la définition du taux de Kendall suivante :

$$\tau_i = \sum_{\substack{j,j' \\ j<j'}} (P_{ijj'} - Q_{ijj'}) \; / \; \binom{d}{2}$$

On a

$$p^o = \frac{1}{n} \sum_i \sum_{\substack{j,j' \\ j<j'}} P_{ijj'} \; / \; \binom{d}{2}$$

D'autre part, on note

$$q^o = \frac{1}{n} \sum_i \sum_{\substack{j,j' \\ j<j'}} Q_{ijj'} \; / \; \binom{d}{2}$$

q^o s'interprète comme une proportion de désaccord.

On prend la définition du tau de Kendall moyen $\bar{\tau}$:

$$\bar{\tau} = p^o - q^o$$

Il est clair que $\bar{\tau} = 2\,p^o - 1$

On voit que $\bar{\tau}$ s'apparente à l'indice d'accord majoritaire (MAI) calculé sur un tableau de données dichotomiques $nx\binom{d}{2}$ (Voir, par exemple, Landis et Koch, 1975).

Exprimé suivant cette dernière forme, $\bar{\tau}$, comme p^o , s'interprète donc comme un barycentre : α_k pour p^o devient $2\alpha_k - 1$ pour $\bar{\tau}$. Les poids sont les mêmes.

Remarque
========

Parmi les derniers travaux concernant la mesure Kappa, on peut citer Verducci, Mack et DeGroot (1988). Ils étudient un estimateur du maximum de vraisemblance de la mesure Kappa, étant donné, sur le sujet i, un vecteur aléatoire d'évaluations $(X_{i1}, \ldots, X_{id})$, où chaque composante ne peut prendre que deux valeurs possibles (on a des données dichotomiques), 0 ou 1. Pour ce faire, différentes hypothèses sur les X_{ij} les conduisent à considérer différentes distributions sur la variable $Y_i = \sum_j X_{ij}$ (la plus simple étant évidemment la distribution binominale lorsque les X_{ij} , j = 1, ..., d, sont supposées indépendantes et identiquement distribuées). ■

CHAPITRE 2

Une Approche par la Théorie des Partitions

On se place à partir de maintenant dans le cadre d'une variable qualitative nominale à valeurs dans $(e_1, \ldots, e_L)$.

Landis et Koch (1975) emploient l'expression "Pairwise agreement considerations" lorsqu'ils présentent les différentes mesures basées sur l'analyse des couples d'évaluations et a fortiori, sur des études de corrélation.

Parmi ces mesures, la mesure Kappa pondérée est essentiellement une mesure interobservateurs faisant intervenir des comparaisons 2 à 2 comme le souligne Schouten (1980).

On voudrait maintenant analyser l'accord interobservateurs de façon globale sans faire intervenir la notion de couple d'observateurs ou d'évaluations.

2.1 NOTION DE FONCTION D'INDICE

En s'inspirant de la construction de p^o ou du tau de Kendall, on envisage le degré de l'accord sur le sujet i comme une "fonction d'indice" φ_i définie sur l'ensemble des partitions de d :

$$D = (p_1, \ldots, p_\delta)$$

et à valeurs réelles ; et une mesure du degré d'accord sera la valeur de φ_i correspondant à la partition observée.

Exemple

On considère la fonction d'indice :

$$\varphi_i : D \longrightarrow [0,1]$$
$$(n_{ie})_e \longmapsto \sum_e \binom{n_{ie}}{2} \Big/ \binom{d}{2}$$

Une mesure du degré d'accord est alors :

$$\varphi_i((n_{ie})_e)$$

qui est une estimation de Π_{ee} lorsque le sujet i seul est observé. ■

Sur l'ensemble des sujets, on définit le degré d'accord φ comme un barycentre des φ_i étant donnée une distribution de

poids sur (1, ..., n) :

φ est une application de D^n sur $\mathbb{R}$.

En accordant le poids $\frac{1}{n}$ à chacun des sujets, on a :

$$\varphi = \frac{1}{n} \Sigma \varphi_i$$

On suppose, comme on a fait jusqu'à présent, que chaque sujet est évalué par le même nombre d'observateurs sélectionnés dans une population d'observateurs.

Ainsi, on admet momentanément que les fonctions φ_i sont toutes identiques à une même fonction :

$$\psi : D \longrightarrow \mathbb{R}$$

Soit $p_k \in D$:

$$\forall i \in (1, \ldots, n),\ \psi(p_k) = \varphi_i(p_k) = \alpha_k$$

($\alpha_k \in \mathbb{R}$).

Soit $p^{(n)} = (p_{m_i}) \in D^n$ une suite de partitions de d comprenant n termes, autant que de sujets.

Si l'on reprend, dans ce cadre général, des notations déjà employées, on pose :

$$\lambda_k = \text{Card } (i\ ,\ 1 \leq i \leq n \mid p_{m_i} = p_k\ ,\ p_k \in D) \quad ,\ 1 \leq k \leq \delta$$

Définition

La mesure du degré d'accord entre les observateurs est la suivante :

$$\varphi(p^{(n)}) = \frac{1}{n} \sum_{i=1}^{n} \varphi_i(p_{m_i}) = \frac{1}{n} \sum_{1}^{\delta} \lambda_k \alpha_k \qquad \blacksquare$$

Remarque

Une telle mesure, en présence d'une variable qualitative nominale, ne dépend pas, comme cela s'impose, du codage défini sur les catégories, mais seulement de la répartition des observateurs en groupes d'observateurs. A l'intérieur de chacun de ces groupes, les observateurs sont en accord 2 à 2. L'existence d'un groupe unique est équivalent à l'existence de l'accord parfait. $\blacksquare$

On a ainsi dégagé l'importance de la suite $(\alpha_k)_{k=1,\ldots\delta}$ dans la définition du degré d'accord.

Exemple

Fixons d = 8 .

D'où, δ = Card D = 22 .

Soit $(n_e)_e$, une partition de D : la fonction d'indice ψ considérée est celle que l'on a déjà rencontrée ; elle est telle que :

$$\psi((n_e)_e) = \sum_e \binom{n_e}{2} / \binom{d}{2}$$

Le tableau suivant nous donne les différentes valeurs possibles de ψ :

p_k	$(n_e)_e$	α_k
p_1	(8)	1
p_2	(7 1)	0,75
p_3	(6 2)	0,57
p_4	(6 1 1)	0,536
p_5	(5 3)	0,464
p_6	(5 2 1)	0,39
p_7	(5 1 1 1)	0,357
p_8	(4 4)	0,429
p_9	(4 3 1)	0,32
p_{10}	(4 2 2)	0,286
p_{11}	(4 2 1 1)	0,25
p_{12}	(4 1 1 1 1)	0,214
p_{13}	(3 3 2)	0,25
p_{14}	(3 3 1 1)	0,214
p_{15}	(3 2 2 1)	0,179
p_{16}	(3 2 1 1 1)	0,14
p_{17}	(3 1 1 1 1 1)	0,107
p_{18}	(2 2 2 2)	0,14
p_{19}	(2 2 2 1 1)	0,107
p_{20}	(2 2 1 1 1 1)	0,07
p_{21}	(2 1 1 1 1 1 1)	0,036
p_{22}	(1 1 1 1 1 1 1 1)	0

Cet exemple montre que la mesure Kappa, construite sur les différentes valeurs possibles de p^o , lesquelles sont fonctions des α_k donc des p_k , comporte un certain arbitraire :

Sur un sujet particulier, $p^o = \sum_e \binom{n_e}{2} / \binom{d}{2}$ ($(n_e)_e$ est une partition de d) prend la valeur $\frac{1}{4}$ aussi bien lorsqu'on observe p_{11} que p_{13} . Lorsqu'on observe p_8 , p^o prend une valeur 0,429 supérieure à celle qu'elle prend lorsqu'on observe p_7 , soit 0,357.

Il est un fait que la suite (α_k) doit être déterminée en tenant compte d'une idée a priori dévelopée sur la notion de degré d'accord et en conséquence elle doit être déterminée au vu d'un ordre sur D et en fait à l'aide d'une fonction ψ qui respecte cet ordre. Cette idée a priori peut être exprimée à l'aide d'un certain nombre de critères ainsi que le montreront les exemples abordés plus loin.

2.2 ORDRE SUR L'ENSEMBLE D DES PARTITIONS DE L'ENTIER d

2.2.1 Exemple d'un ordre sur D

Complétons les d-m termes d'une partition de d ($0 \leq m \leq d-1$) par un nombre m de zéros de façon à se ramener à un vecteur de $\mathbb{R}^d$ à d composantes.

On forme ainsi un ensemble $D_o = \{v_1 , \ldots , v_\delta\}$ d'éléments de $\mathbb{R}^d$ sur lesquels on considère l'ordre (Marshall et Olkin, 1979) :

Etant donné v_1 , $v_2 \in D_o$, $v_1=(w^1_1 , \ldots, w^1_d)'$ et, $v_2 = (w^2_1 , \ldots, w^2_d)'$,

$$v_1 <_o v_2 \iff \begin{cases} \forall q , 1\leq q\leq d , \sum_{q}^{d} w^1_{(i)} \leq \sum_{q}^{d} w^2_{(i)} \\ \sum_{1}^{d} w^1_{(i)} = \sum_{1}^{d} w^2_{(i)} \end{cases}$$

(v_1 est majoré par v_2).

Les éléments de D et de D_o sont en bijection par construction. Supposons que les notations associent $p_k \in D$ à $v_k \in D_o$ (par exemple, si $p_k = (3\ 1)$, $v_k = (3,1,0,0)'$).

On peut définir l'ordre partiel sur D :

$$p_k \underset{\sim}{\leq} p_{k'} \iff v_k <_o v_{k'} , v_k \in D_o , v_{k'} \in D_o$$

Soit D_o^M un sous-ensemble de cardinal maximal de D_o , sur lequel l'ordre partiel $\underset{\sim}{\leq}$ sur D devient un ordre total.

Modifions l'ensemble D_o en un ensemble D'_o obtenu en remplaçant tous les éléments de $D_o \backslash D_o^M$ par $S=(1, \ldots, 1)'$.

On définit alors un ordre total <' sur D par :

$$p_k <' p_{k'} \iff v_k <_o v_{k'} , v_k \in D'_o , v_{k'} \in D'_o$$

Interprétation

Les partitions de d sont ordonnées en fonction d'un "regroupement progressif" des observateurs tenant compte de la décroissance du nombre de groupes.
Lorsque cette décroissance est interrompue, une partition pourra être assimilable à S qui traduit une "dispersion" maximale des évaluations.

Dans l'exemple précédent, on pourra donc assimiler p_5 , p_8 , p_{13} et p_{18} à p_{22} , puisqu'il existe pour chacune de ces partitions, une partition qui ne lui est pas comparable au sens de $\underset{\sim}{\leq}$. ■

2.2.2 Une caractérisation générale d'un ordre sur D

Même si l'on peut trouver une interprétation globale d'une relation d'ordre sur D , il n'est pas toujours aisé d'interpréter localement une relation du type $p_k<p_{k'}$.

De plus, le nombre de comparaisons possibles des partitions de d 2 à 2 , devient vite très grand et donc aussi le nombre d'interprétations locales d'une relation du type $p_k<p_{k'}$, puisque asymptotiquement,

$$\delta \cong \frac{1}{4d\sqrt{3}} \exp \pi\sqrt{2d/3}$$

(Andrews, 1976).

C'est pourquoi on a recours au procédé suivant :

Soit une suite $(\beta_k)_{k=1, \ldots, \delta}$ de réels tels que $\beta_{(1)} \neq \beta_{(\delta)}$.

On considère l'intervalle :

$$[\beta_{(1)} , \beta_{(\delta)}]$$

et une partition de cet intervalle en s ($s \geq 2$) intervalles contigus :

$$C_1 = [\beta_{(1)} , b_1[$$
$$\vdots$$
$$C_r = [b_{r-1} , b_r[\qquad (s>2 \ ; \ 2 \leq r \leq s-1)$$
$$\vdots$$
$$C_s = [b_{s-1} , \beta_{(\delta)}]$$

($b_r \in \mathbb{R}$) . Et on pourra poser $b_s = \beta_{(\delta)}$.

Ceci permet de définir une fonction en escalier :

$$g : [\beta_{(1)} , \beta_{(\delta)}] \longrightarrow \mathbb{R}$$

à l'aide d'une suite réelle $(\gamma_r)_{r=1, \ldots, s}$

par : $x \in C_r \Rightarrow g(x) = \gamma_r$

Par exemple, on prendra : $\gamma_r = \frac{r-1}{s-1}$.

Un ordre < sur D sera alors défini par :

$$p_k < p_{k'} \iff g(\beta_k) \leq g(\beta_{k'})$$

Cette construction formelle d'un ordre sur D permet d'établir, en fixant s inférieur à δ , des équivalences entre les partitions et ceci diminue le nombre d'interprétations locales d'une relation du type $p_k < p_{k'}$, susceptibles d'être effectuées.

Exemple

Considérons l'ordre défini précédemment :

$$p_k <' p_{k'} \iff v_k <_o v_{k'} \ , \ v_k \in D'_o \ , \ v_{k'} \in D'_o$$

Rangeons les partitions par ordre décroissant. Posons ensuite

$$v_k = (w^k_1 , \ldots , w^k_d) ,$$

et

$$\beta_k = \frac{1}{d-1} \sum_{q>1} \sum_{i=q}^{d} w^k_{(i)}$$

Prenons $\forall r , r'$, diam C_r = diam $C_{r'}$; s = 11 ; d'où

$$g(C_r) = \frac{1}{10} (r-1)$$

On peut alors comparer l'ordre <' avec l'ordre <" défini par :

$$p_k <" p_{k'} \iff g(\beta_k) \leq g(\beta_{k'})$$

grâce au tableau (d=8) :

p_k	ordre pour <'	β_k	$g(\beta_k)$	ordre pour <"
p_1	1	8	1	1
p_2	2	7,86	1	2
p_3	3	7,7	1	3
p_4	4	7,57	0,9	4
p_5	6	7,57	0,9	5
p_6	7	7,43	0,9	6
p_7	9	7,14	0,8	8
p_8	10	7,43	0,9	9
p_9	11	7,29	0,9	7
p_{10}	12	7,14	0,8	10
p_{11}	14	7	0,8	11
p_{12}	15	6,86	0,7	13
p_{13}	16	7	0,8	12
p_{14}	17	6,86	0,7	14
p_{15}	19	6,7	0,7	15
p_{16}	20	6,43	0,6	16
p_{17}	21	5,86	0,5	18
p_{18}	5	6,29	0,6	17
p_{19}	8	6,14	0,5	19
p_{20}	13	5,7	0,4	20
p_{21}	18	5	0,3	21
p_{22}	22	4	0	22

(Les accolades indiquent des regroupements d'indices de partitions équivalentes).

On constate que p_5 est une partition "de tête" pour <" et une partition "de queue" pour <' .

On constate aussi que <" ne fait pas beaucoup bouger p_8 , p_{13} et p_{18} contrairement à <' . ■

2.3. ORDRE SUR D DEFINI PAR UNE FONCTION D'INDICE

Une fonction d'indice ψ définie sur chaque sujet peut induire un ordre < sur D dans la mesure où la suite (β_k) est déterminée et où l'on pose :

$$\psi(p_k) = \alpha_k = g(\beta_k)$$

Ainsi on pourra écrire :

$$p_k < p_{k'} \iff \psi(p_k) \leq \psi(p_{k'})$$

De cette manière et d'après une idée a priori sur la notion de degré d'accord, il reste maintenant à déterminer des critères de construction des suites (β_k) et (b_r) , k=1, ..., δ ; r=1, ... , s .

CHAPITRE 3

Variance d'une Partition et Indice d'Accord

Dans le chapitre précédent, on s'est intéressé à déterminer des relations d'ordre sur les partitions de l'entier d .

En lien avec une analyse de concordance, on vient de définir :

$$\begin{cases} \psi(p_k) = g(\beta_k) \\ p_k < p_{k'} \iff \psi(p_k) \leq \psi(p_{k'}) \end{cases}$$

$\psi(p_k)$ étant une mesure du degré d'accord entre les observateurs lorsque l'on dispose de la donnée $p_k \in D$.

On propose dorénavant de prendre g croissante.

Ainsi, g n'apparaît être qu'une fonction destinée à "corriger" la répartition des β_k sur $[\ \beta_{(1)}\ ,\ \beta_{(\delta)}\ ]$ en respectant l'ordre initial sur les β_k .
On précisera plus loin comment choisir g .

3.1. VARIANCE D'UNE PARTITION DE L'ENTIER d

On voit que se fixer des critères pour définir les valeurs α_k de la fonction d'indice ψ revient à se fixer les mêmes critères pour définir les β_k puisque α_k est simplement liée à β_k par une fonction en escalier croissante g .

On prend les notations suivantes :

$$p_k = (z_1 \dots z_m) \in D$$

où $1 \leq z_q \leq d$ pour $1 \leq q \leq m$;
et on rappelle que $\sum_{1}^{m} z_q = d$.

Critères de construction des β_k

Critère 1 :

$\beta_{(\delta)}$ doit être associé à la partition (d) pour laquelle l'accord est parfait. ■

En fonction de ce 1er critère, on donne les deux critères supplémentaires suivants :

Critère 2 :

Les β_k doivent marquer une tendance à décroitre avec m .
m s'interprête en effet comme un certain degré d'"éparpillement" des évaluations, m = 1 étant synonyme de l'accord parfait. ■

Critère 3 :

Les β_k doivent marquer une tendance à croître avec $\sum_q \left(z_q - \frac{d}{m}\right)^2$, $m \geq 2$.

Cette dernière expression représente un écart de la répartition des évaluations décrite en terme d'effectifs par $\begin{pmatrix} z_1 & \cdots & z_m \end{pmatrix}$, à une répartition uniforme fictive décrite par $\begin{pmatrix} \frac{d}{m} & \cdots & \frac{d}{m} \end{pmatrix}$ (m composantes).

Une telle répartition formalise, dans une certaine mesure, la notion de désaccord puisqu'elle s'interprête comme une "neutralisation" des évaluations. ■

En fonction des trois critères énoncés plus haut, on définit la suite β_k ainsi :

si $m \geq 2$, $\beta_k = \frac{1}{m} \sum_{q=1}^{m} \left(z_q - \frac{d}{m}\right)^2$

si $m = 1$, $p_k = (d)$. On suppose par exemple que $p_1 = (d)$.

On prendra pour β_1 un majorant quelconque de $\left\{\beta_k \mid p_k \in D \setminus \{(d)\}\right\}$ tel que $\forall k > 1$, $\beta_k < \beta_1$.

On voit que, lorsque $m \geq 2$, chacun des deux facteurs, $\frac{1}{m}$ et $\Sigma\left(z_q - \frac{d}{m}\right)^2$,imprime à β_k la tendance escomptée dans la réponse à chacun des critères 2 et 3 .

On définit ainsi β_k , pour $m \geq 2$,comme une variance

empirique des composantes des p_k puisque l'espérance mathématique empirique de ces composantes est alors égale à

$$\frac{1}{m} \Sigma z_q = \frac{d}{m}$$

On écrira : Var $p_k = \frac{1}{m} \Sigma \left(z_q - \frac{d}{m}\right)^2$.

Les propositions suivantes sont autant de justifications au choix que l'on a fait de la suite (β_k) d'après les trois critères que l'on s'est fixés :

Proposition 1 (Une justification liée au 1er critère)

Supposons que $\forall p = (z_1 \ \cdots \ z_m) \in D$, on ait $(z_1 \ \cdots \ z_m) = (z_{(1)} \ \cdots \ z_{(m)})$.

Dans ces conditions, le problème

$$\max_{p \in D} \text{Var } p$$

possède une unique solution en (1 d-1).

En effet,

On note : $p_2 = (1\ d-1)$

et $D^* = D \setminus \{(d)\}$

On montre la propriété $\mathbb{P}$ suivante :

$$\forall p^* \in D^* \setminus \{p_2\} \ , \ \text{Var } p^* < \text{Var } p_2$$

Démonstration par récurrence sur d .

$\mathbb{P}$ est vraie pour d = 2 .

Supposons $\mathbb{P}$ vraie pour 2, ..., d-1 .

Démontrons que $\mathbb{P}$ est alors vraie pour d ($d \geq 3$).

Une partition de D^* sera :

ou bien : cas 1 : (w d-w) , $w \in \{1,\ldots,d-1\}$

ou bien : cas 2 : $(w\ z_1 \ldots z_m)$ avec :

$$\begin{cases} 2 \leq m \leq d-w \\ w \in \{1,\ldots,d-2\} \\ \sum_1^m z_q = d-w \end{cases}$$

Cas 1

La variance de (w d-w) vaut

$$\frac{d^2}{4} - wd + w^2$$

Les valeurs de cette forme quadratique en w sont d'autant plus grandes que w est petit ou grand, c'est-à-dire, w = 1 ou w = d-1

Cas 2

La variance de $(w\ z_1 \ldots z_m)$ vaut

$$\frac{1}{m+1} w^2 + \frac{m}{m+1} \left[\frac{1}{m} \sum z_q^2 - \frac{1}{m^2}(d-w)^2 \right] + \frac{(d-w)^2}{m(m+1)} - \frac{d^2}{(m+1)^2}$$

c'est-à-dire :

$$\frac{(mw+w-d)^2}{m(m+1)^2} + \frac{m}{m+1} \left[\frac{1}{m} \sum z_q^2 - \frac{1}{m^2}(d-w)^2 \right]$$

D'après l'hypothèse de récurrence, il faut montrer que l'on a l'inégalité :

$$\frac{(mw+w-d)^2}{m(m+1)^2} + \frac{m}{m+1} \left[\frac{(d-w)^2}{4} - d + w + 1 \right] < \frac{d^2}{4} - d + 1$$

c'est-à-dire :

$$(c) \quad \frac{m^2+4m+4}{4m(m+1)} w^2 - \left[\frac{d(m^2+4)-2m^2}{2m(m+1)} \right] w + \frac{d^2}{m(m+1)^2} - \frac{d^2}{4(m+1)} + \frac{d-1}{m+1} < 0$$

où l'expression de gauche est une forme quadratique en w.

Cette dernière inégalité sera vérifiée si donc elle l'est pour w = 1 et w = d-1 .

a) w = 1

- On vérifie aisément que si m=2, 3 ou 4, l'inégalité (c) est vraie et donc $\mathbb{P}$ est vraie.
- Si $m \geq 5$ (et donc $d \geq 6$), l'inégalité (c) qui est équivalente à (en multipliant par 4m(m+1)) :

$$(5-2d)m^2 - d(d-4)m - 8d + 4 + \frac{4d^2}{m+1} < 0$$

est vraie si :

$$-5d^2 + 4md - 8d + 4 + \frac{2}{3}d^2 < 0$$

et a fortiori, puisque $m \leq d-1$,si :

$$-\frac{1}{3}d^2 - 12d + 4 < 0$$

Donc $\mathbb{P}$ est encore vraie.

b) $w = d-1$

En multipliant (c) par $4m(m+1)$, (c) devient :

$$(d-2)[-m^2+4m] + 4(d+2)\left[-1+\frac{1}{m+1}\right] - (d-2)m + \frac{8}{(d-2)(m+1)} < 0$$

- Lorsque $m \geq 4$, $\frac{8}{(d-2)(m+1)} \leq \frac{8}{5}$ et $(d-2)m \geq 4$; de plus les deux crochets sont négatifs.

- Lorsque $m = 2$ ou 3, l'inégalité est vraie.

Donc $\mathbb{P}$ est vraie.

Ainsi, $\mathbb{P}$ est toujours vraie. ■

Propositions 2 (Une justification liée au 2e critère)

Soit $p^* = (z_1 \ldots z_{q'} \ldots z_m) \in D^*$; $1 \leq q' \leq m$; $2 \leq z_{q'} \leq d-1$.
On note p^{**} la partition : $(z_1 \ldots z_{q'} - w \ldots z_m\ w)$
telle que $1 \leq w < z_{q'}$.
Alors on a :

$$m \geq -\frac{1}{2} + \frac{1}{2}\sqrt{1+\frac{2d^2}{(z_{q'}-w)w}} \Longrightarrow \text{Var } p^{**} < \text{Var } p^* .$$

En effet,

$$\text{Var } p^{**} = \frac{1}{m+1}\left[\sum_{q \neq q'} z_q^2 + (z_{q'}-w)^2 + w^2\right] - \frac{d^2}{(m+1)^2}$$

$$= \frac{m}{m+1}\left\{\frac{1}{m}\sum_q z_q^2 - \frac{d^2}{m^2}\right\} + \frac{d^2}{m(m+1)} - \frac{d^2}{(m+1)^2} - \frac{2(z_{q'}-w)w}{m+1}$$

Pour avoir $\text{Var } p^{**} < \text{Var } p^*$, il suffit donc que :

$$\frac{d^2}{m(m+1)} - \frac{d^2}{(m+1)^2} - \frac{2(z_{q'}-w)w}{m+1} \leq 0$$

Et cette inégalité sera vérifiée à l'extérieur de l'intervalle délimité par les racines du polynôme en m :

$$(m+1)d^2 - md^2 - 2m(m+1)(z_{q'}-w)w$$

Une condition suffisante pour avoir
$\text{Var } p^{**} < \text{Var } p^*$ est donc :

$$m \geq -\frac{1}{2} + \frac{1}{2}\sqrt{1+\frac{2d^2}{(z_{q'}-w)w}}$$ ■

Proposition 3 (Une justification liée au 3e critère)

Soit $p^* = (z_1 \ldots z_{q'} \ldots z_{q''} \ldots z_m) \in D^*$,
$1 \leq q' < q'' \leq m$; $2 \leq z_{q'} \leq z_{q''} \leq d-2$.

Soit p^{**} la partition : $(z_1 \ldots z_{q'}-1 \ldots z_{q''}+1 \ldots z_m)$.
Alors on a : $\text{Var } p^* < \text{Var } p^{**}$.

En effet,

$$\text{Var } p^* = \frac{1}{m} \sum_q z_q^2 - \frac{d^2}{m^2}$$

$$\text{Var } p^{**} = \frac{1}{m} \sum_{\substack{q \neq q' \\ q \neq q''}} z_q^2 + \frac{(z_{q'} - 1)^2 + (z_{q''} + 1)^2}{m} - \frac{d^2}{m^2}$$

D'où :

$$\text{Var } p^{**} - \text{Var } p^* = \frac{2(z_{q''} - z_{q'})}{m} + \frac{2}{m} > 0 \qquad \blacksquare$$

Le tableau ci-dessous nous montre, pour d=8 , l'ordre 0 sur D^* induit par la suite $(\beta_k) = (\text{Var } p_k)$.

Il permet également de comparer 0 avec l'ordre 0_1 (resp. 0_2) induit par la suite $\left(\sum (z_q - \frac{d}{m})^2 \right)$ (resp. la suite $(\frac{1}{m})$) (chacun des termes généraux de ces deux dernières suites étant bien sûr associé par une relation du type $\beta_k = \frac{1}{m} \sum (z_q - \frac{d}{m})^2$) .

p_k	$\sum (z_q - \frac{d}{m})^2$	0_1	$\frac{1}{m}$	0_2	β_k	ordre 0 sur D^*
p_1			1			
p_2	18	1	0,5	1	9	1
p_3	8	5	0,5	2	4	3
p_4	16,67	2	0,33	4	5,56	2
p_5	2	13	0,5	3	1	9
p_6	8,67	4	0,33	5	2,89	5
p_7	12	3	0,25	9	3	4
p_8	0		0,5		0	
p_9	4,67	8	0,33	6	1,56	6
p_{10}	2,67	12	0,33	7	0,89	11
p_{11}	6	7	0,25	10	1,5	7
p_{12}	7,2	6	0,2	13	1,44	8
p_{13}	0,67	18	0,33	8	0,22	16
p_{14}	4	9	0,25	11	1	10
p_{15}	2	14	0,25	12	0,5	14
p_{16}	3,2	11	0,2	14	0,64	12
p_{17}	3,33	10	0,17	16	0,56	13
p_{18}	0		0,25		0	
p_{19}	1,2	16	0,2	15	0,24	15
p_{20}	1,33	15	0,17	17	0,22	17
p_{21}	0,86	17	0,14	18	0,12	18
p_{22}	0		0,125		0	

On a donc montré précédemment que la suite $(\text{Var } p_k)$ possède un plus grand élément égal à $(\frac{d}{2} - 1)^2$ (Proposition 1) :

$$\forall k \in \{ 1, \ldots, \delta \} \quad , \quad \text{Var } p_k \leq (\frac{d}{2} - 1)^2$$

D'autre part, posons, de nouveau, $p_1 = (d)$.

Il est clair que l'on peut minorer la suite $(\text{Var } p_k)_{k=2,\ldots,\delta}$ par 0 :

$$\forall k \in \{ 2,\ldots, \delta \} , \qquad 0 \leq \text{Var } p_k$$

Un tel minorant n'est pas toujours un plus petit élément. On le voit dans quelques cas, comme par exemple :

1) $L = 2$; d , impair

2) $L < d$; d , premier

D'autre part, dans l'exemple précédent, on constate que

$$\forall k \in \{ 2, \ldots, \delta \} \quad , \text{Var } p_k \in [0,9]$$

et l'on remarque que moins de 20 % des valeurs de $(\text{Var } p_k)_{k=2, \ldots, 22}$ sont supérieures ou égales à 3:

$$\text{Card } \{k \mid 2 \leq k \leq 22 \ ; \text{Var } p_k \in [3,9]\} = 4$$

Voilà pourquoi on n'utilise pas directement la suite (β_k) pour mesurer l'accord, mais une suite (α_k) déduite de (β_k) en transformant les différents termes de (β_k) à l'aide d'une fonction en escalier g .

3.2. REPARTITION DES β_k SUR $[\beta_{(1)}, \beta_{(\delta)}]$

On peut donner une idée de la répartition des β_k sur $[\beta_{(1)}, \beta_{(\delta)}]$ à l'aide de quelques estimations.

Or, on a déjà indiqué que l'existence d'un grand nombre d'éléments appartenant à D, pour d "grand", rendait complexe l'étude de cet ensemble : Pour les 100 premières valeurs de d, on a dans Andrews (1976) les valeurs correspondantes de δ. Donnons un extrait de cette correspondance :

d	1	2	3	8	13	18	23	100
δ	1	2	3	22	101	385	1255	$\simeq 19.10^7$

Ainsi, si l'on s'intéresse, par exemple, à la proportion de partitions de D^* dont la variance est supérieure à une valeur donnée (ainsi qu'on l'a fait pour d=8), on est amené à estimer cette proportion de la manière qui va suivre.

D'abord quelques notations :

Soit V_m la variance d'un élément de D^* comportant m composantes ($m \geq 2$) noté $(z_1 \ldots z_m)$ $(\sum_q z_q = d)$.

Soit z un élément de la suite $(z_q)_q$. On peut toujours

supposer sans perdre de généralité que $z = z_1$.

On note V^z_{m-1} la variance de la partition de l'entier d-z , $(z_2 \dots z_m)$ et l'on renote V_m , V^z_m .

Lemme 1 : $V^z_m = \frac{m-1}{m} V^z_{m-1} + \frac{1}{m-1} (z-\frac{d}{m})^2$ ■

Lemme 2 : On a l'encadrement de V^z_m suivant :

$$\frac{1}{m-1} (z-\frac{d}{m})^2 \leq V^z_m \leq \frac{m-1}{m} (\frac{d-z}{2}-1)^2 + \frac{1}{m-1} (z-\frac{d}{m})^2$$

En effet,

il suffit, d'après le lemme 1, d'encadrer V^z_m à l'aide d'un encadrement de V^z_{m-1} . ■

Une estimation moyenne de V^z_m , soit V_A , peut alors être obtenue à partir de la moyenne des deux bornes définies au lemme 2 :

$$V_A = \frac{m-1}{2m} (\frac{d-z}{2}-1)^2 + \frac{1}{m-1} (z-\frac{d}{m})^2$$

D'où :

$$V_A = \frac{1}{2}(1-\frac{1}{m})(\frac{d-z}{2}-1)^2 + \frac{z^2}{m-1} - 2zd(\frac{1}{m-1}-\frac{1}{m}) + d^2(\frac{1}{m-1}-\frac{1}{m}-\frac{1}{m^2})$$

Une telle estimation dépend de m et de z.

On préfère, par souci de simplification, d'après ce qu'on a déjà indiqué, trouver maintenant une estimation de V^z_m indépendante de m.

Pour cela, on remarque que

$$2 \leq m \leq d-z+1$$

Toute valeur de $\frac{1}{m}$ peut donc être approchée par la moyenne suivante :

$$\frac{1}{d-z} \sum_{2}^{d-z+1} \frac{1}{m} = \frac{1}{d-z} [\gamma - 1 + \log (d-z+1) + 0((d-z+1)^{-1})]$$

où γ est la constante d'Euler : $\gamma \simeq .577$.

De même, toute valeur de $\frac{1}{m^2}$ peut être approchée par la moyenne suivante :

$$\frac{1}{d-z} \sum_{2}^{d-z+1} \frac{1}{m^2} = \frac{1}{d-z} [\frac{\pi^2}{6} - 1 - \frac{1}{d-z+1} + 0((d-z+1)^{-2})]$$

(Voir, par exemple, Ellison et Mendès-France, 1975).

Ainsi, V_A pourra être approchée, pour des valeurs suffisamment élevées de d-z, par V_z :

$$V_z = \left\{ \frac{1}{2} - \frac{1}{2(d-z)} [\gamma - 1 + \log (d-z+1)] \right\} (\frac{d-z}{2} - 1)^2$$

$$+ [\gamma + \log (d-z)] \frac{z^2}{d-z} + (2-\frac{\pi^2}{6}) \frac{d^2}{d-z} - \frac{2zd}{d-z+1}$$

V_z est une estimation de V^z_m indépendante de m .

Considérons z tel que :

$$\frac{d}{2} \leq z \leq d-3$$

Ceci implique $z = z_{(m)}$ et $d-z \geq 3$.
Par exemple, on pourra prendre $z = [\frac{2}{3}d]$ pour $d \geq 7$.

Notons, pour plus de précision, $\delta(d)$, le nombre de partitions de l'entier d.

Lemme 3

Soit $\frac{d}{2} \leq z \leq d-1$.

Le nombre d'éléments de la suite (Var p_k) ($p_k \in D^*$) tels que la plus grande composante de p_k soit supérieure ou égale à z, est égal à

$$\sum_{k=1}^{d-z} \delta(k)$$

En effet,
la valeur de z est unique parmi les composantes de la partition. ■

D'autre part, notons $(v_h^z)_h$ la sous-suite de la suite $(\text{Var } p_k)_k$ telle que la plus grande composante de p_k soit z.
V_z représente une valeur moyenne des éléments de $(v_h^z)_h$.
Une estimation du nombre δ_z^o d'éléments de la suite (Var p_k) ($p_k \in D^*$) supérieurs à V_z ($\frac{d}{2} \leq z \leq d-1$) est donc égale,

d'après le lemme 3, à :

$$\delta_z = \sum_{k=1}^{d-z+1} \delta(k) + \frac{1}{2}\,\delta(d-z)$$

En effet, on peut estimer le nombre d'éléments de la suite $(v_h^z)_h$ supérieurs à V_z à la moitié du nombre d'éléments de cette suite lequel vaut $\delta(d-z)$.

On pourrait en fait montrer, d'après le type d'approximation réalisé sur $\frac{1}{m}$ et $\frac{1}{m^2}$, que

$$\delta_z^1 = \sum_{k=1}^{d-z} \delta(k)$$

est une "meilleure" estimation de δ_z^o que ne l'est δ_z .

De sorte que l'on peut écrire, en référence au résultat obtenu pour d=8 :

$$\delta_z^o = \text{Card}\ \{\ k \mid 2 \leq k \leq \delta\ ;\ \text{Var } p_k \in [V_z\ ,\ (\tfrac{d}{2}-1)^2]\} \simeq \delta_z^1$$

$(\frac{d}{2} \leq z \leq d-1)$

Soit $z = [\frac{2}{3}d]$: cette valeur correspond à l'existence d'un groupe d'observateurs en accord entre eux, constitué de près des 2/3 des observateurs.

On considère t_1 , le pourcentage estimé d'éléments de

(Var p_k) supérieurs à V_z :

$$t_1 = 100\ \delta_z^1 / \delta(d)$$

On considère aussi t_2 , le taux d'occupation sur $[0, (\frac{d}{2}-1)^2]$ des éléments de (Var p_k) supérieurs à V_z :

$$t_2 = 100\ [1 - V_z / (\frac{d}{2}-1)^2]$$

Le tableau suivant permet d'apprécier la répartition des Var p_k (c'est-à-dire des β_k) sur l'intervalle $[\beta_{(1)} , \beta_{(\delta)}]$:

d	$\delta(d)$	$z=[\frac{2}{3}d]$	V_z	$(\frac{d}{2}-1)^2$	δ_z	δ_z^o	δ_z^1	t_1	t_2
7	15	4	0,82	6,25	4,5	6	6	40	86,9
8	22	5	1,62	9	4,5	6	6	27,3	82
9	30	6	2,777	12,25	4,5	5	6	20	77,3
10	42	6	2,896	16	8,5	10	11	26,2	81,9
13	101	8	6,136	30,25	14,5	19	18	17,8	79,7
18	385	12	15,8	64	23,5		29	7,5	75,3
23	1255	15	25,02	110,25	55		66	5,3	77,3

3.3. CATEGORISATION DES β_k : CONSTRUCTION DE (b_r)

On retient le critère de construction de (b_r) suivant :

$\forall r , r'$, le nombre d'éléments distincts de (β_k) dans C_r est égal au nombre d'éléments distincts de (β_k) dans $C_{r'}$.

On voit immédiatement que les β_k ne pourront généralement satisfaire au critère que de manière approchée puisqu'il s'agit d'un calcul en nombres entiers.

Cependant, on considère les exceptions suivantes :

(i) $[b_{s-1} , b_s]$ ne contient que β_1.
Il faut donc nécessairement avoir :

$$(\frac{d}{2} - 1)^2 < b_{s-1} \leq \beta_1 \leq b_s$$

Ceci revient à isoler le cas de l'accord parfait.

(ii) Si d est "grand", on n'impose que le critère soit satisfait que pour une sous-suite des (β_k) seulement :

Par exemple on considère les β_k tels que les partitions p_k associées aient une plus grande composante z vérifiant :

$$z \geq \frac{d}{2}$$

$\frac{d}{2}$ représente le seuil de l'existence d'un groupe majoritaire d'observateurs en accord entre eux.
Ainsi, il suffira de construire la suite (b_r) au vu d'un nombre restreint de partitions égal à

$$\sum_{1}^{[d/2]} \delta(k)$$

Pratiquement

Il semble naturel de prendre :

$s = \delta$ pour $d = 2, 3, 5$

$s = 4$ pour $d = 4$

$s = 9$ pour $d = 6$

$s = 15$ pour $d = 7$

$s = 20$ pour $d = 8$

Ainsi, le critère pourra être satisfait avec un élément de (β_k) distinct des autres dans chaque sous-intervalle.

Pour $d \geq 9$, on propose de prendre $s = 21$.
Ainsi, les valeurs de [0,1] prises par les γ_r auront un écart minimum de .05 .

En ce qui concerne les b_r , on prend :

$$b_{s-1} = b_s = \beta_1 = \left(\frac{d}{2}-1\right)^2 + \varepsilon$$

($\varepsilon > 0$ arbitraire ; par exemple : $\varepsilon = 1$)

La détermination des b_r pour $r = 2, \ldots, s-2$ ($s \geq 4$) pourra être obtenue, par exemple, par la méthode de la cumulée inverse (Diday, 1982) lorsque $d \geq 9$.

Exemple

$d = 3$; $\gamma_r = \frac{r-1}{s-1}$; $b_{s-1} = b_s = \beta_1 = \left(\frac{d}{2}-1\right)^2 + 1$; on a :

p_k	Var p_k	$\beta_{(1)} = 0$	
(1 1 1) (2 1) (3)	0 0.25 0	$b_1 = 0.1$ $b_2 = 1.25$ $b_3 = 1.25$	$\gamma_1 = 0$ $\gamma_2 = 0.5$ $\gamma_3 = 1$

(b_1 est arbitraire sur l'intervalle $]0, \frac{1}{4}[$).

Ainsi :

$\beta_{(1)} = 0 \in C_1$; $\beta_{(2)} = .25 \in C_2$; $\beta_{(3)} = \beta_1 = 1.25 \in C_3$.

Ces trois valeurs sont réparties uniformement sur [0, 1] à l'aide de la fonction g qui les transforme en $\gamma_1 = 0$; $\gamma_2 = .5$ et $\gamma_3 = 1$.

3.4 INDICE D'ACCORD

On vient de déterminer les sous-intervalles C_1, C_2, ..., C_s permettant de catégoriser β_1, β_2, ..., β_δ .

Ainsi, pour $1 \leq r \leq s-1$, chaque sous-intervalle C_r caractérise une famille de degrés d'accord correspondant à une sous-suite de la suite (Var p_k) . C_s caractérise l'accord parfait.

L'observation d'un certain nombre d'éléments de (Var p_k) dans

une expérience d'évaluation de n sujets, conduit à l'observation d'une suite $p^{(n)} = (p_{m_i})$ de partitions (suivant les notations déjà employées).

On pose :

Si $r \leq s-1$, $\mu_r = \text{Card}\ \{i,\ 1 \leq i \leq n \mid \text{Var}\ p_{m_i} \in C_r\}$

et $\mu_s = \text{Card}\ \{i,\ 1 \leq i \leq n \mid p_{m_i} = (d)\}$

(μ_s est le nombre d'observations de l'accord parfait).

L'indice d'accord J proposé est la mesure du degré d'accord définie page 115, à savoir $\varphi(p^{(n)}) = (1/n) \sum_1^{\delta} \lambda_k g(\beta_k)$, où l'on choisit $\gamma_r = \frac{r-1}{s-1}$ comme étant les différentes valeurs de g :

$$J = \frac{1}{n} \sum_1^s \mu_r \frac{r-1}{s-1}$$

On a :

$$\sum_1^s \frac{\mu_r}{n} = 1$$

CHAPITRE 4

Corrections, Estimations, Tests et Observation Spatiale

4.1. CORRECTIONS

On peut envisager d'apporter une correction aux β_k , $k = 2, \ldots, \delta$ dans le but,

- soit d'améliorer la définition d'un ordre sur les partitions en liaison avec la notion d'accord : on parlera de correction globale.
- soit de distinguer, dans l'ensemble des évaluations, celles qui apportent une forte contribution à la valeur globale de l'indice J. On admettra que de telles évaluations ont pu être facilement conceptualisées par les observateurs. On parlera de correction locale.

4.1.1. Correction globale et modèle de concentration

On note de nouveau $p_k = (z_1 \ldots z_m)$, un élément quelconque

de D à m composantes ($m \geq 2$) .

On aura deux groupes majoritaires d'observateurs qui s'opposeront dans la mesure où l'on pourra écrire :

$$\frac{z_{(m-1)}}{z_{(m)}} = 1$$

β_k pourra donc avoir une tendance à décroître avec $\frac{z_{(m-1)}}{z_{(m)}}$ et l'on pourra envisager de prendre comme valeur pour β_k , non pas Var p_k , mais :

$$\beta_k = \left\{1 - \left[\frac{z_{(m-1)}}{z_{(m)}}\right]^2\right\} \text{ Var } p_k$$

$k = 2, \ldots, \delta$.

Ainsi, la fonction qui, à $\frac{z_{(m-1)}}{z_{(m)}}$, associe $\left\{1 - \left[\frac{z_{(m-1)}}{z_{(m)}}\right]^2\right\}$, est strictement convexe sur [0, 1] et n'affecte que faiblement Var p_k lorsque $z_{(m-1)}$ est proche de 1, c'est-à-dire $z_{(m)}$ proche de d-1.

Remarque
========

Une telle correction ne modifie pas l'ordre des valeurs prises par les β_k sur l'ensemble des p_k possédant des composantes toutes égales.

En effet,

dans un tel cas, on a, en particulier $z_{(m-1)} = z_{(m)}$

et donc simultanément :

$$1 - \left[\frac{z_{(m-1)}}{z_{(m)}}\right]^2 = 0 \text{ et Var } p_k = 0 .$$

Ce sont alors deux valeurs redondantes . ■

La remarque précédente ouvre donc la voie à une modification éventuelle du facteur Var p_k : on peut envisager un nouveau facteur répondant aux critères 1 et 2 et ne répondant plus au critère 3 (page 126), mais à un nouveau critère 3', comme on va le voir maintenant.

On voudrait associer la notion d'accord avec la notion de concentration de même qu'on a associé, au chapitre 1, la notion d'accord avec la notion de corrélation.

On suppose toujours que $p_k = (z_1 \ldots z_m)$ et l'on convient de poser si $m > 1$:

$$z_{m+1} = \ldots = z_d = 0$$

Critère 3'

Les β_k doivent marquer une tendance à croître avec $\sum_{q=1}^{d} z_q^2$ ($\sum_{q=1}^{d} z_q = d$) qui est une fonction convexe des z_q possédant un minimum global sur les entiers strictement positifs en $z_q = 1$ pour $q = 1, \ldots, d$ ("éparpillement des

évaluations") et un maximum global sur les entiers positifs ou nuls en $z_{q_o} = d$, $q_o \in (1, \ldots, d)$ et $z_q = 0$ pour $q \neq q_o$ (accord parfait). ■

$\sum_1^d (z_q / \sum_1^d z_q)^2$ est strictement Schur-Convexe (Marshall et Olkin, 1979).

On a :

$$d \leq \sum_1^d z_q^2 \leq d^2$$

On remarque que $\sum_{q=1}^{d} z_q^2$ est l'un des termes qui interviennent dans l'expression $\sum_{q=1}^{m} (z_q - \frac{d}{m})^2$ puisque :

$$\sum_{q=1}^{m} (z_q - \frac{d}{m})^2 = \sum_{q=1}^{d} z_q^2 - \frac{d^2}{m}$$

Ainsi, le critère 3 concerne un moment centré des z_q et le critère 3' concerne un moment non centré des z_q .

UN MODELE DE CONCENTRATION

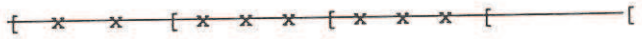
Tout groupe d'observateurs en accord entre eux va être considéré comme contribuant au désaccord dans la mesure où ce groupe ne sera pas unique.

Un tel groupe d'observateurs d'effectif z_q (z_q<d) va être caractérisé par un certain nombre de points égal à $d-z_q$ sur une maille-1 d'une grille à d mailles sur un segment de la droite réelle.
L'occupation d'une maille-1 va correspondre de façon biunivoque à l'existence d'un groupe d'observateurs en accord entre eux.

Le désaccord, provoqué par l'existence d'un tel groupe et associé à la notion de non-concentration, va être traduit par une dispersion uniforme des $d-z_q$ points sur la maille-1 correspondant au groupe en question.

Exemple

d = 4
La partition (2 1 1) est représentée par le schéma :



A chacun de ces points d'une maille-1 on attribue une pondération égale à $z_q/d(d-z_q)$ ce qui introduit une généralisation dans la construction du coefficient de concentration $D(\underline{Y})$ que l'on va renoter ici D_k puisqu'il est calculé au vu de p_k.
En effet, les notations actuelles permettent d'écrire, de

manière simplifiée,

$$n_{m+1} = \sum_{u=1}^{d^m} \frac{K_{u,m}}{d} N_{u,m}$$ (voir page 72)

et dans cette formule, les pondérations sur les points correspondent au facteur 1/d.

La somme des pondérations est, bien entendu, encore égale à 1.

Ces pondérations sont croissantes avec z_q.

En se référant au chapitre 3 de la 1ère partie, il est alors clair que pour la partition $p_k = (z_1 \dots z_m)$, on a :

- $K_{1,0} = d$; $N_{1,0} = d-m$
- $K_{q,1} = d-z_q$; $N_{q,1} = z_q$
- $K_{u,m} = 1$; $N_{u,m} = d-1$ pour $m \geq 2$

D'où le coefficient de concentration D_k vaut :

$$D_k = \frac{d-m}{d} + \frac{1}{d^2} \sum_q (d-z_q) \frac{z_q}{d(d-z_q)} z_q + \sum_3^\infty \frac{d-1}{d^m}$$

$$= -\frac{m}{d} + \frac{1}{d^3} \sum_{q=1}^{m} z_q^2 + 1 + \frac{1}{d^2}$$

On remarque que, par construction, D_k est calculé lorsque

l'on n'est pas dans la situation de l'accord parfait mais ceci conduit à redéfinir β_k , $1 \leq k \leq \delta$, comme étant égal à D_k , en étendant les valeurs de D_k au cas où $\sum z_q^2 = d^2$ et $m = 1$ qui correspond précisément à l'accord parfait.

Et ainsi, les critères 1, 2 et 3' sont satisfaits. β_1 est alors égal à $1 + 1/d^2$.
On peut également prendre une valeur de β_k corrigée :

$$\beta_k = \left\{1 - \left[\frac{z_{(m-1)}}{z_{(m)}}\right]^2\right\} D_k$$

Remarque
========

D_k détermine, en soi, un indice de concordance défini sur $[0, 1+1/d^2]$, d'autant plus petit, en cas d'accord parfait, que d est plus grand. ■

4.1.2 Corrections locales : indice intraclasse et indice de substituabilité

UN INDICE INTRACLASSE D'ACCORD

La notion d'indice intraclasse est dégagée, d'un point de vue probabiliste, par Schouten (1980) à l'aide de la mesure Kappa.

L'idée est la suivante :

On sélectionne une évaluation e dans l'ensemble $\{e_1, \ldots, e_L\}$.

Si un sujet n'a pas reçu l'évaluation e, on le supprime de l'expérience.

L'expérience va donc porter sur un nombre n' restreint de sujets.

De plus, sur un sujet donné, la partition de d observée : $(z_1 \ldots z_e \ldots z_m)$ va comporter une composante non nulle z_e égale au nombre de fois que l'évaluation e aura été portée sur ce sujet.

Dans ce contexte, définir une mesure de l'accord conditionnée par une présence fréquente de l'évaluation e, sur chacun des n' sujets, c'est définir un indice intraclasse concernant e.

Un tel indice pourra ainsi être construit en imposant à β_k une tendance à croître avec $\frac{z_e - 1}{z_{(m)} - 1}$ (lorsque $z_{(m)} \neq 1$).

On prendra comme valeur pour β_k :

$$\beta_k = \left[\frac{z_e - 1}{z_{(m)} - 1}\right]^2 \text{Var } p_k$$

$k = 2, \ldots, \delta$. Et on conviendra que $\beta_k = 0$, si $z_{(m)} = 1$.

Ainsi, la fonction qui, à $\frac{z_e - 1}{z_{(m)} - 1}$, associe $\left[\frac{z_e - 1}{z_{(m)} - 1}\right]^2$, est strictement concave sur $[0,1]$ et n'affecte que faiblement Var p_k lorsque z_e est proche de $z_{(m)}$.

On ne modifie pas la suite (b_r) et l'indice a une expression du type $(1/n') \sum_1^s \mu'_r \gamma_r$.

On remarque que si $\left[\frac{z_e - 1}{z_{(m)} - 1}\right]^2$ est proche de 0, on ne peut rien dire quant à l'accord global.

Il pourra donc être intéressant d'étendre les valeurs possibles de l'indice intraclasse sur $[-1, 1]$ en considérant comme facteur correcteur non pas $\left[\frac{z_e - 1}{z_{(m)} - 1}\right]^2$, mais l'expression :

$$\left[\frac{z_e - 1}{z_{(m)} - 1}\right]^2 + \left[\frac{1}{2} \frac{z_e - (d/2) + (1/4)}{\sqrt{(z_e - (d/2) + (1/4))^2}} - \frac{1}{2}\right]$$

où le crochet vaut

0 si z_e est grand ($z_e \geq d/2$)

-1 si z_e est petit ($z_e < d/2$)

De sorte que si le 1er terme de cette expression est proche de 0, le 2ème est proche de -1 et ainsi, en valeur absolue, cette expression n'affectera que faiblement Var p_k.

A un grand écart entre z_e et $z_{(m)}$ et, conjointement, à une grande valeur de $z_{(m)}$, ("fort" accord) va correspondre une grande valeur négative de Var p_k.

On pourra dire, dans ce cas, que l'accord est obtenu "malgré" la présence de l'évaluation e.

UN INDICE DE SUBSTITUABILITE

Considérons deux évaluations, e et e' .

On peut s'intéresser à l'indice intraclasse concernant e' , noté $J_{e'}$, calculé sur l'ensemble restreint des n" sujets (n" > 0) ayant reçu conjointement les évaluations e et e' .

Si $J_{e'}$ est élevé, on peut parler d'un accord obtenu "grâce à" e' et "malgré" e : e' est alors une évaluation substituable à e .

On admettra une telle substituabilité si n" est au moins égal à la moitié de n' (on rappelle que n' est le nombre de sujets ayant reçu l'évaluation e).

Ainsi, la valeur de β_k envisagée pour l'indice de substituabilité $J_{e'}$, substituant e' à e, est :

$$\beta_k = \left[\frac{z_{e'}-1}{z_{(m)}-1}\right]^2 \left[\frac{1}{2} \frac{2n"-n'+1/4}{\sqrt{(2n"-n'+1/4)^2}} + \frac{1}{2}\right] \text{Var } p_k$$

(le crochet vaut 1 si n" ≥ n'/2 et 0 sinon).

On ne modifie pas la suite (b_r) et l'indice de substituabilité est du type (1/n") $\sum_1^s \mu_r" \gamma_r$. Si $J_{e'}$ est faible, on ne parlera pas de substituabilité de e' par e mais de compatibilité de e' et de e .

4.2 ESTIMATIONS ET TESTS

Que ce soit pour l'un ou l'autre des indices considérés plus haut, on constate que la méthode d'analyse de concordance développée ici, par la théorie des partitions, nous met en présence d'un suite $(\mu_1, \ldots, \mu_s)$ telle que $\sum_1^s \mu_r = n$.

Cette suite représente un résumé de l'ensemble des données initiales .

Dans un tel cadre, observer l'accord parfait quelque soit le sujet évalué, c'est observer :

$$\mu_s = n$$

(et dans ce cas, $\mu_1 = \ldots = \mu_{s-1} = 0$) .

On a alors

$$\mu_s = n \Longrightarrow \mu_s = \mu_{(s)}$$

Ceci conduit à considérer $(\mu_1 , \ldots, \mu_s)$ comme étant une réalisation d'une variable multinomiale $(F_1 , \ldots, F_s)$ paramétrée par $(n, s, f_1 , \ldots, f_s)$ avec $\sum_1^s f_r = 1$:

n, s et d sont fixés ; les f_r sont inconnus ; les différents échantillons possibles correspondent à différents sujets possibles dont le nombre et toujours égal à n.

On s'intéresse ensuite à la statistique d'ordre :

$$F_{max} = \max (F_1 , \ldots, F_s) = F_{(s)} .$$

On a les deux propriétés suivantes :

a) $\frac{n}{s} \leq F_{max} \leq n$

b) $P(F_{max} > \mu) = 1 - P(F_1<\mu, \ldots, F_s<\mu) \quad (\mu \in \mathbb{R})$

A priori on ne connaît pas les paramètres f_r ; or ce sont leurs valeurs qui amèneraient à définir une mesure exacte du degré d'accord qui serait un accord entre les observateurs envisagé quelque soit l'ensemble de sujets examiné .

Plusieurs méthodes d'analyse des f_r peuvent être considérées.

4.2.1 Estimations

Johnson et Kotz (1969) nous donnent une estimation simultanée par intervalle de confiance pour chaque f_r au seuil de $1-\alpha$; les bornes d'un tel intervalle sont :

$$\frac{\chi^2_{s-1,1-\alpha} + 2\mu_r \pm \{\chi^2_{s-1,1-\alpha}[\chi^2_{s-1,1-\alpha} + 4\mu_r n^{-1}(n-\mu_r)]\}^{1/2}}{2(n + \chi^2_{s-1,1-\alpha})}$$

Les expressions précédentes fournissent un intervalle de confiance pour l'indice global.

4.2.2 Tests

Pour savoir empiriquement quels sont ceux parmi les f_r que l'on peut considérer comme les plus grands, on fait l'hypothèse H_o suivante :

$$H_o : f_1 = \ldots = f_s$$

Contre l'hypothèse alternative :

$$H_r : \forall r' \in (1, \ldots, s) , r' \neq r , f_r > f_{r'}$$

On utilise la statistique F_{max} , de réalisation μ_r , comme variable discriminante en déterminant la valeur μ_o vérifiant :

$$P\ (F_{max} > \mu_o | H_o) = \alpha$$

(α, erreur de 1ère espèce).

En cas de rejet de H_o , c'est-à-dire si l'on observe $\mu_r > \mu_o$, on admettra que la valeur du paramètre f_r est significativement élevée et contribuera à indiquer quel type de partition caractérise la concordance.

Les s tests réalisés séparent les s paramètres en deux classes de paramètres, ceux pour lesquels on aura rejeté H_o (classe I) et ceux pour lesquels on aura accepté H_o (classe II).

Pour admettre que l'accord entre les observateurs est significatif, on peut se définir un ou plusieurs critères :

On retiendra les deux critères A et B suivants devant être vérifiés simultanément :

Critère A sur la composition de la classe I :

Si f_r appartient à la classe II, on demande que l'une des trois éventualités suivantes au moins soit vérifiée :

(i) $r = s$

(ii) $\forall r'$, $r' \leq r$, $f_{r'}$ appartient à la classe II

(iii) f_{r+1} appartient à la classe I

Interprétation :

On n'exige pas que l'accord parfait soit fréquent ; tout type de désaccord doit être peu fréquent ; enfin, à tout type d'accord peu fréquent, on doit pouvoir faire correspondre un "meilleur" type d'accord fréquent. ■

Critère B sur l'importance réciproque des classes I et II :

Si $\{f_{r_1}, \ldots, f_{r_u}\}$, $1 \leq r_t \leq s$, $1 \leq t \leq u$, est l'ensemble des éléments de la classe I, on pose :

$$f_I = \sum_{t=1}^{u} f_{r_t}$$

On considère la variable binomiale $F_I\ (n, f_I)$ de réalisation $\sum_{t=1}^{u} \mu_{r_t}$.

Le test de l'hypothèse $f_I = \frac{1}{2}$ contre $f_I > \frac{1}{2}$, basé sur la région critique $(\mu_I , \ldots, n)$, défini par

$$P(F_I > \mu_I) = \alpha$$

(α, erreur de première espèce), doit conduire à rejeter l'hypothèse.

Interprétation :

Un tel test donne une indication concernant l'aspect prépondérant de l'accord sur le désaccord dans la mesure où le critère A est vérifié . ■

Réalisation pratique du test de H_o contre H_r

Mallows (1968) démontre que l'expression $P(F_{max} > \mu_o | H_o)$ possède l'encadrement suivant :

$$1 - e^{-n(1-F(\mu_o))} \leq P(F_{max} > \mu_o | H_o) \leq n(1-F(\mu_o))$$

où F est la fonction de répartition d'une variable binomiale de paramètres n et $\frac{1}{s}$.

Si l'on calcule maintenant la valeur μ qui vérifie :

$$1 - e^{-n(1-F(\mu))} = \alpha$$

la zone de rejet de H_o définie par $(\mu, \ldots, n)$ est alors d'amplitude inférieure à celle que définit l'équation $P(F_{max} > \mu_o | H_o) = \alpha$.

On conçoit ainsi que le test basé sur $(\mu, \ldots, n)$ est "plus exigeant" que celui de niveau α et que, pour ce test, les calculs ont l'avantage d'être simplifiés.

L'encadrement précédent permet en fait de définir une frontière pour la région critique, non réduite à un point, pour un α donné.

4.3 OBSERVATION SPATIALE

On peut s'intéresser, dans cette partie, à la comparaison des d observateurs pour savoir quels sont ceux qui "influencent" l'accord global mesuré par l'indice J, et quels sont ceux qui "influencent" le désaccord global (on parlera d'observateurs déviants), au sens suivant :

J est calculé pour un ensemble de d observateurs.

Notons R l'un de ces observateurs.
Considérons l'ensemble des observateurs auquel on retranche R et notons J_R l'indice de l'accord calculé pour ce nouvel ensemble.

Considérons de même, R', différent de R, et $J_{R'}$.

On pourra comparer les influences de R et de R' sur l'accord global, en regardant si J_R est plus petit ou plus grand que $J_{R'}$.

C'est une approche employée dans diverses études empiriques et en particulier celle réalisée par Hermann et Hovaguimian (1983). Ces auteurs parlent, à ce propos, de biais entre les observateurs.
Une comparaison des observateurs de type géométrique consisterait à employer la méthode graphique que l'on va définir

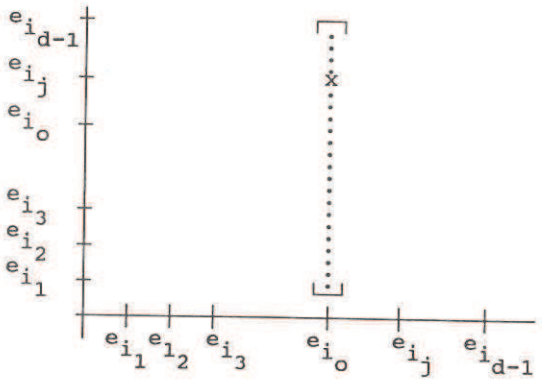
maintenant.

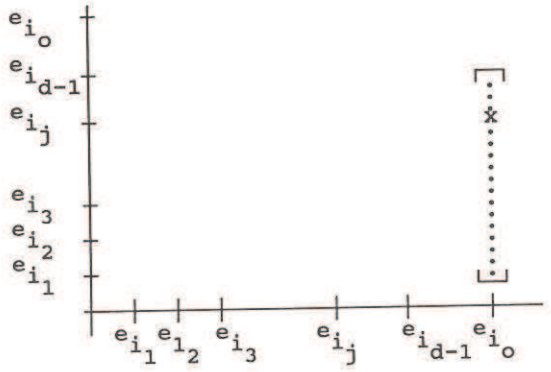
Soit, dans le plan, un repère rectangulaire. Sur chacun des deux axes la coordonnée d'un point sera obtenue au vu d'un ensemble de graduations régulièrement espacées et associées à une échelle nominale.

Précisons à quoi correspond l'ensemble des graduations :
Considérons l'observateur R : on suppose qu'il a attribué au sujet i l'évaluation e_{i_o} .

Supposons que l'un des d-1 autres observateurs a attribué, à ce même sujet i, l'évaluation e_{i_j} où e_{i_j} est un élément d'une suite d'évaluations notée $(e_{i_j})_{j=1, \ldots, d-1}$.

Une association immédiate entre les graduations et les évaluations permet de reporter dans le plan d-1 points du type (e_{i_o}, e_{i_j}) : 2 cas de figure :





Ces points se situent sur un segment de droite vertical d'extrémités (e_{i_o}, e_{i_1}) et $(e_{i_o}, e_{i_{d-1}})$.

L'accord de R avec l'un des autres observateurs, ayant porté l'évaluation e_{i_j} sur le sujet i, se traduit par $i_o = i_j$, ce qui situe le point (e_{i_o}, e_{i_j}) sur la diagonale du repère.

On reporte ainsi dans le plan n(d-1) points en faisant varier i de 1 à n (n sujets) et l'on constitue de la sorte, pour l'observateur R, un nuage de points Ω_R dans le plan.

Remarque 1
==========

Si l'accord est parfait quelque soit le sujet, Ω_R est inclu dans la diagonale du repère. ■

Remarque 2
==========

Il n'est pas nécessaire, pour définir Ω_R , d'introduire sur chacun des axes, des graduations associées à des éléments de $(e_1, \ldots, e_L)$ n'ayant fait l'objet d'aucune attribution de la part des observateurs au cours de l'expérience. C'est d'ailleurs ce qui nous est suggéré par le graphique ci-dessus où les évaluations e_{i_j} et e_{i_o} ont effectivement été attribuées par les observateurs. ■

On peut construire, de cette manière, d nuages du type Ω_R , chacun de ces nuages caractérisant le comportement d'un observateur particulier vis-à-vis de l'ensemble des autres observateurs.

L'étude de chacun de ces nuages par une analyse de forme telle qu'elle est décrite à la fin de la première partie, à l'aide du coefficient de concentration (p.85), conduit à la construction de d lignes polygonales que l'on peut comparer entre elles dans $\mathbb{R}^2$.

Il est clair qu'en cas d'accord parfait, quelque soit le sujet, les d lignes polygonales seront alignées sur la diagonale du repère.

Une ligne polygonale "éloignée" de cette diagonale correspondra à un observateur dit déviant.

Les d déviations des lignes polygonales à la diagonale pourront être mesurées par d surfaces, chacune étant calculée entre la ligne polygonale correspondante et la diagonale.

* * *

CONCLUSION

Une nouvelle approche de l'analyse d'un nuage de points dans R^p a conduit à observer un tel nuage du point de vue de la concentration de ses points grâce à une méthode de réseaux.

Dans une perspective de "Projection Pursuit" le nouvel indice de concentration a pu être comparé aux indices déjà utilisés dans la littérature.

Différentes qualités de cet indice ont été mises en valeur de par sa construction même et ses propriétés. La notion de concentration a été ainsi confrontée à celle de distance interpoints et d'entropie.

Il apparaît particulièrement intéressant de décrire un nuage par la concentration de tout ou partie de l'ensemble de ses points lorsqu'il n'est pas opportun de postuler, au départ, une structure déterminée sur le nuage : On abandonne entre autre, toute référence à une structure normale.

De cette manière, on n'a plus besoin, par exemple, de justifier une inertie dans un espace euclidien, par le fait qu'une norme peut fournir, à une homothétie 1/n près, dans un modèle normal, une estimation efficace de la variance des composantes d'une projection du nuage.

On se dégage également de la notion d'entropie définie comme un écart à une distribution a priori sur un nuage de points.

De plus, on a pu vérifier que la concentration était une mesure peu sensible à l'existence de points aberrants.

Précisons que l'interprétation quantitative de l'indice de concentration est aisée aux valeurs extrêmes. En revanche, en ce qui concerne les valeurs intermédiaires, elles sont comparables entre elles sur l'échelle des valeurs possibles de l'indice. Ceci permet d'analyser les différentes concentrations d'un nuage en projection, en observant sur différentes projections les phases monotones des valeurs correspondantes de l'indice.

Ces considérations nous amènent à mentionner l'aspect algorithmique du problème de l'optimisation de l'indice sous la contrainte de l'orthogonalité de la projection.

Il a été possible d'illustrer le problème de l'optimisation dans R^2 en passant d'un problème contraint à un problème non contraint aisément résoluble graphiquement.

En revanche, en dimension p, p>2, la résolution numérique du problème d'optimisation de l'indice passe par un programme d'optimisation performant vu la taille du problème ; une telle résolution constitue un objectif à venir tout à fait intéressant.

La précision des résultats doit, en particulier, être inversement proportionnelle à p . Cette précision est essentiellement fonction du type de troncature effectué sur le développement en série qui définit l'indice de concentration.

Le modèle de concentration appliqué en analyse de concordance, apporte une sorte de justification à l'utilisation de la variance d'une partition comme indice de classement des éléments de l'ensemble des partitions possibles de d observateurs évaluant un sujet donné.

En effet, concentration et variance, dans le cadre de la théorie des partitions, apparaissent dépendre d'une même fonction Shur-convexe.

A ce propos, les propriétés de la variance d'une partition, qui sont tout à fait exceptionnelles, méritent d'être vérifiées dans le cas de la concentration.

Dans cette perspective de la concordance, la concentration trouve un début de généralisation dans le calcul de l'indice pour un ensemble de points possédant des multiplicités fractionnaires.

Il apparaît enfin important pour de futurs développements de mentionner la possibilité d'introduire une matrice de

proximité sur les différentes valeurs possibles de la variable qualitative nominale considérée.

On peut prévoir qu'avec une telle matrice, l'indice de concordance intraclasse qui mesure l'accord sur chaque modalité de la variable, prendra des valeurs plus significatives qu'en l'absence d'une telle matrice, au sens du test paramétrique que l'on a construit.

- ANNEXE 1 -

RESEAU LINEAIRE SUR UN INTERVALLE BORNE, SEMI-OUVERT A DROITE

De La Vallée Poussin (1950) utilise les méthodes de réseaux dans R^p, pour certains problèmes topologiques de recouvrements et pour définir des dérivées de fonctions multidimensionnelles. Plus tard, C., Tricot (1971 et 1988) utilisera les réseaux dans l'étude de la concentration d'ensembles bornés, en relation avec la densité définie sur de tels ensembles.

En dimension 1, un réseau, sur un intervalle $[a,b[$, semi-ouvert à droite, est défini par une suite de grilles, $(G_m)_{m=0,1..}$, où G_m est construite de la manière suivante :

- $G_o = [a,b[$
- Construction de G_1 : On recouvre $[a,b[$ par une suite finie d'intervalles contigus, semi-ouverts à droite, de même diamètre, en plaçant sur $[a,b]$ des points a_i régulièrement espacés, appelés noeuds, tels que : $a = a_o < a_1 < \ldots < a_n = b$, s'il y a n+1 noeuds.
  L'ensemble $\{[a_k , a_{k+1}[, k=0, \ldots, n-1\}$ définit la grille G_1 .
- Construction de G_2 : G_2 est la réunion de n grilles, chacune étant définie sur un élément particulier (chaque fois différent) de G_1 , à l'aide d'un certain nombre de noeuds, de la même manière qui a permis de définir G_1 .
  G_2 est donc obtenue en ajoutant aux noeuds de la grille G_1 de nouveaux noeuds, et en formant l'ensemble correspondant d'intervalles contigus, semi-ouverts à droite auxquels on impose d'avoir même diamètre.
- Construction de G_m : De manière récurrente, G_m est définie à partir de G_{m-1} en ajoutant aux noeuds de la grille G_{m-1} de nouveaux noeuds.

Les intervalles semi-ouverts à droite, d'extrémités deux noeuds consécutifs de l'ensemble des noeuds de G_m , doivent avoir même diamètre. On les appelle les mailles-m de G_m .

Un réseau sur $[a,b[$, dépend du mode de génération des noeuds d'une grille à l'autre.

On conviendra d'ajouter n-1 noeuds à l'intérieur de chaque maille-(m-1) de la grille G_{m-1} pour constituer l'ensemble des noeuds de la grille G_m , m=1,2,... . Ainsi,

- G_o est définie à l'aide de deux noeuds : a et b
- G_1 est définie à l'aide de n+1 noeuds : a, b et n-1 points de $]a,b[$
- G_2 est définie à l'aide de $n+1 + n(n-1) = n^2+1$ noeuds

. G_m est définie à l'aide de $n^{m-1} + 1 + n^{m-1}(n-1) = n^m + 1$ noeuds

Le nombre de maille-m de la grille G_m est n^m.

* * *

ANACONDA POUR 4 OBSERVATEURS (VARIABLE QUAL NOM)

RESUME DU PROGRAMME : ANACONDA.FOR

```
C """"""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""
C
C ENTREE DE LA TAILLE DES DONNEES ET DES TRIPLETS
C OBSERVATEUR-SUJET-EVALUATION, EN FORMAT LIBRE, DANS UN FICHIER
C
C
C ECRITURE DE LA MATRICE DES TRIPLETS OBSERV.-SUJET-EVALUAT.
C
C
C CALCUL DU NOMBRE D'OBSERVATEURS : variable NBCL
C CALCUL DU NOMBRE DE SUJETS PAR OBSERVATEUR : variable NH
C
C
C CALCUL DU NOMBRE DE SUJETS : variable NBPT
C CALCUL DES NUMEROS DE SUJET : variable NXX
C CALCUL DU NOMBRE D'OBSERVATEURS PAR SUJET : variable NR
C
C
C ECRITURE DE NBCL, NBPT, et de LA TAILLE DES DONNEES
C
C
C ECRITURE DES NUMEROS DE SUJET ET DU NOMBRE D'EVALUAT. PAR SUJET
C
C
C ECRITURE DES NUMEROS D'OBSERVATEUR
C
C
C LES COORDONNEES DES POINTS DU NUAGE DE CHAQUE OBSERVATEUR
C
C
C ECRITURE DU NUMERO DE L'OBSERV. ET DES COORD. DES PTS DE SON
C NUAGE
C
C
C ECRITURE DU NOMBRE D'EVALUATIONS POSEES PAR CHAQUE OBSERVATEUR
C
C
C ECRITURE DES DIFFERENTES EVALUATIONS POSEES
C
```

```
c
c ECRITURE DU NOMBRE D'EVALUATIONS DIFFERENTES POSEES
c
c
c CALCUL DES PARTITIONS POUR L'ENSEMBLE DES SUJETS
c
c
c ECRITURE DES n PARTITIONS ET DES EVALUATIONS
c ASSOCIES A LEURS DIFFERENTES COMPOSANTES
c
c
c         LES 5 INDICES DE CLASSEMENT DES PARTITIONS CONSIDERES:
c            _ LA VARIANCE   LA VARIANCE CORRIGEE
c            _ LA CONCENTRATION _ LA CONCENTRATION CORRIGEE
c            _ LE PAR (le coefficient de cartwright)
c
c CALCUL DES INDICES DE CLASSEMENT POUR LES n SUJETS
c
c
c CALCUL DES 5 INDICES DE CLASSEMENT THEORIQUES DES PARTITIONS
c LORSQUE d=4
c
c
c ECRITURE DES PARTITIONS ET DE LEURS INDICES DE CLASSEMENT
c (les partitions sont ordonnees d'apres leurs variances)
c
c
c CALCUL DES PROBABILITES BINOMIALES (n, 1/4) ET (n, 1/2)
c ET DES PROBABILITES BINOMIALES CUMULEES ASSOCIEES
c
c
c ECRITURE DES TABLES DE PROB.  BINOMIALES (n, 1/4) ET (n, 1/2)
c
c
c           INDICE DE CLASSEMENT : LA VARIANCE
c            --------------------   ===========
c
c ECRITURE DES VALEURS DE L'INDICE DE CLASSEMENT POUR LES n SUJETS
c
c
c CALCUL DES FREQUENCES DE PARTITIONS TYPES SUR L'ENSEMBLE DES
c SUJETS
c
c
c ECRITURE DES VALEURS TYPES DE LA VARIANCE
c ET LEUR FREQUENCE (non nulle)
c
c TESTS SUR LA STATISTIQUE FMAX; LES RESULTATS
c
c
c TESTS SUR LA STATISTIQUE BINOMIALE (n, 1/2); LES RESULTATS
c
c
c           INDICE DE CLASSEMENT : LA VARIANCE CORRIGEE
c           --------------------   ====================
c
c ECRITURE DES VALEURS DE L'INDICE DE CLASSEMENT POUR LES n SUJETS
c
```

```
c
c CALCUL DES FREQUENCES DE PARTITIONS TYPES SUR L'ENSEMBLE DES
c SUJETS
c
c
c ECRITURE DES VALEURS TYPES DE LA VARIANCE CORRIGEE
c ET LEUR FREQUENCE (non nulle)
c
c
c TESTS SUR LA STATISTIQUE FMAX; LES RESULTATS
c
c
c TESTS SUR LA STATISTIQUE BINOMIALE (n, 1/2); LES RESULTATS
c
c           INDICE DE CLASSEMENT : LA CONCENTRATION
c           --------------------   ================
c
c ECRITURE DES VALEURS DE L'INDICE DE CLASSEMENT POUR LES n SUJETS
c
c
c CALCUL DES FREQUENCES DE PARTITIONS TYPES SUR L'ENSEMBLE DES
c SUJETS
c
c
c ECRITURE DES VALEURS TYPES DE LA CONCENTRATION
c ET LEUR FREQUENCE (non nulle)
c
c
c TESTS SUR LA STATISTIQUE FMAX; LES RESULTATS
c
c
c TESTS SUR LA STATISTIQUE BINOMIALE (n, 1/2); LES RESULTATS
c
c
c           INDICE DE CLASSEMENT : LA CONCENTRATION CORRIGEE
c           --------------------   ========================
c
c ECRITURE DES VALEURS DE L'INDICE DE CLASSEMENT POUR LES n SUJETS
c
c
c CALCUL DES FREQUENCES DE PARTITIONS TYPES SUR L'ENSEMBLE DES
c SUJETS
c
c
c ECRITURE DES VALEURS TYPES DE LA CONCENTRATION CORRIGEE
c ET LEUR FREQUENCE (non nulle)
c
c
c TESTS SUR LA STATISTIQUE FMAX; LES RESULTATS
c
c
c TESTS SUR LA STATISTIQUE BINOMIALE (n, 1/2); LES RESULTATS
c
c
c           INDICE DE CLASSEMENT : LE PAR
c           --------------------   ======
c
c ECRITURE DES VALEURS DE L'INDICE DE CLASSEMENT POUR LES n SUJETS
c
```

```
c
c CALCUL DES FREQUENCES DE PARTITIONS TYPES SUR L'ENSEMBLE DES
c SUJETS
c
c ECRITURE DES VALEURS TYPES DU PAR ET LEUR FREQUENCE (non nulle)
c
c
c CALCULS DES COEFFICIENTS D'ACCORD
c
c
c SORTIE DES RESULTATS DANS UN FICHIER
c
c
          stop
          end
c
c
c CINQ SUBROUTINES DE CLASSEMENT
c
c
c SUBROUTINE DE CALCUL DES COMPOSANTES D'UNE PARTITION
c
c
c LES SUBROUTINES DE CALCUL DES CINQ INDICES DE CLASSEMENT DES
c PARTITIONS
c
c """""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""
```

A 2

PROGRAMME ANACONDA.FOR , ECRIT EN FORTRAN 77

```
                character*100 z9(10)
                real xdb(900),xniveau(900),xnotE(920),xnotEE(100)
     .  ,xv1(920),xv2(920),xv3(920),xv4(920),xv5(920)
     .  ,va(10),vaco(10),dk(10),dkco(10),cartw(10),x9(200),x10(200)
                integer*4 nidld(3155),ld1(3155)
     .  ,nc1(920),nc2(3155),nc3(920),nc4(920),nc5(920)
     .  ,not11(920),not12(920),not13(920),not14(920),not15(920)
     .  ,not22(3155),ndiag(1000,50),mccl(920),mmml(920)
     .  ,nxx(200),nx(920),nr(200),nyy(200),nu(200)
     .  ,ngg(920),ng(920),nhh(200),nh(200)
     .  ,mlf(300),menot(300),mxyz1(600),mxyz2(600)
c
c
        open(file='oms.',unit=7,status='old')
        open(file='res.',unit=12,status='new')
c
c
c ENTREE DE LA TAILLE DES DONNEES ET DES TRIPLETS
c OBSERVATEUR-SUJET-EVALUATION, EN FORMAT LIBRE, DANS LE FICHIER
c
c                              ----- oms.-----
c
c SORTIE DES RESULTATS DANS LE FICHIER
c
c                              ----- res.-----
c
c
c
c ENTREES
c
c    1) La taille des donnees,en format libre, dans le fichier: oms.
c    2) Les triplets obs-suj-eval, en format libre, dans le fichier: oms
.
c
c SORTIES
c
c    1) La matrice des triplets observ.-sujet-evaluat.
c    2) Le nb d'observateurs ; le nb de sujets ; La taille des donnees
c    3) Les numeros de sujet et le nb d'evaluations par sujet
c    4) Les numeros d'observateur
c    5) Les points du nuage de chaque observateur
c    6) Les differentes observations posees
c    7) Les partitions et les evaluations associees a leurs composantes
c    8) Les partitions theoriques et 5 indices de classement
c    9) Des tables de probabilites binomiales
c   10) Tests sur les frequences des partitions types pour les 5 indices

c   11) Les indices de concordance et le coefficient de concentration
c
      write(12,*)' '
      write(12,*)' '
      write(12,*)' '
      write(12,*)' '
      write(12,*)'                ',
     .  '                                                          '
      write(12,*)'                ',
     .  '|                                                         '
      write(12,*)'                ',
     .  '|                                                        |'
      write(12,*)'                ',
     .  '| ANACONDA POUR 4 OBSERVATEURS (VARIABLE QUAL NOM)       |'
      write(12,*)'                ',
     .  '|                                                        |'
      write(12,*)'                ',
     .  '|________________________________________________________|'
      write(12,*)' '
      write(12,*)' '
      write(12,*)' '
      write(12,*)' '
      write(12,*)' '
      write(12,*)' '
c
c ENTREE DE LA TAILLE
c
           read(7,*)kq
c
c TITRES
c
      write(12,*)'   Les donnees:     '
      write(12,*)' '
      write(12,*)'                                Observ.    Sujet       Eva',
     .'luat.'
      write(12,*)' '
c
       do i=1,kq
c
c ENTREE DES TRIPLETS OBSERVATEUR-SUJET-EVALUATION
c
c                    OBSERV.  SUJET  EVALUAT.
         read(7,*)nidld(i),ld1(i),not12(i)
c
c ECRITURE DE LA MATRICE DES TRIPLETS OBSERV.-SUJET-EVALUAT.
c
        write(12,*)'              ',nidld(i),ld1(i),not12(i)
c
         mccl(i)=nidld(i)
c
       enddo
c
c
        call  tert(nidld,ld1,not12,kq)
c
c CALCUL DU NOMBRE D'OBSERVATEURS : variable NBCL
c CALCUL DU NOMBRE DE SUJETS PAR OBSERVATEUR : variable NH
c
c                         initialisations:
c
                          nh(1)=1
                          nhh(1)=1
                          nbcl=1
                          nbpt=1
                          nddif=1
                          nbdiad=1
                          nxx(1)=1
                          nyy(1)=1
                          ngg(1)=1
                          ng(1)=1
                          nr(1)=1
                          nu(1)=1
c
             call prime(mccl,kq)
             call part(nbcl,kq,nhh,nh,mccl)
c
c CALCUL DU NOMBRE DE SUJETS : variable NBPT
c CALCUL DES NUMEROS DE SUJET : variable NXX
c CALCUL DU NOMBRE D'OBSERVATEURS PAR SUJET : variable NR
```

```
          do i=1,kq
                    nc3(i)=nidld(i)
                    nc4(i)=ldl(i)
                    nc5(i)=not12(i)
          enddo
c
              call tert(nc3,nc4,nc5,kq)
              call part(nbpt,kq,nxx,nr,nc4)
c
c
          do i=1,kq
                    mccl(i)=nc3(i)
          enddo
c
c MATRICE n*d DES EVALUATIONS : matrice NDIAG
c
          do i=1,nbpt
                                    ns=0
                                if(i.gt.1) then
                                     do jj=1,i-1
                                       ns=ns+nr(jj)
                                     enddo
                                endif
                             nss=nr(i)
             do j=1,nss
                nc1(j)=nc5(j+ns)
             enddo
                      call  prime(nc1,nss)
             do j=1,nss
                 ndiag(i,j)=nc1(j)
             enddo
          enddo
c
c ECRITURE DE NBCL, NBPT, et de LA TAILLE DES DONNEES
c
       write(12,200)
       write(12,*)' '
       write(12,2000)nbcl,nbpt,kq
       write(12,*)' '
       write(12,*)' '
c
c ECRITURE DES NUMEROS DE SUJET ET DU NOMBRE D'EVALUATIONS PAR SUJET
c
      write(12,*)'       ++++++++++++++++++++++++++++++++++++++++++++++++++++',
     .           '++++++++++++++++++++++'
      write(12,*)'       1ere serie: les num. de pat.; 2eme serie: le',
     .           ' nb de clin. par pat.'
      write(12,*)'       ++++++++++++++++++++++++++++++++++++++++++++++++++++',
     .           '++++++++++++++++++++++'
      write(12,*) (nxx(i),i=1,nbpt)
      write(12,*) ' '
      write(12,*) ' '
      write(12,*) (nr(i),i=1,nbpt)
      write(12,*)'       ++++++++++++++++++++++++++++++++++++++++++++++++++++',
     .           '++++++++++++++++++++++'
      write(12,*)'       ++++++++++++++++++++++++++++++++++++++++++++++++++++',
     .           '++++++++++++++++++++++'
c
c ECRITURE DES NUMEROS D'OBSERVATEUR
c
      write(12,*)' '
```

```
      write(12,*)'Les clin'
      write(12,*)(nhh(i),i=1,nbcl)
      write(12,*)' '
c
c
c
        do i=1,kq
                  mmml(i)=nidld(i)
                  nc5(i)=not12(i)
                  nc2(i)=nc5(i)
        enddo
c
c LES COORDONNEES DES POINTS DU NUAGE DE CHAQUE OBSERVATEUR
c
        do i=1,kq
          mlf(i)=mmml(i)
          menot(i)=nc2(i)
        enddo
      do icl=1,4
             ix=0
             iy=0
        do i=1,nbpt
                                    ns=0
                                if(i.gt.1) then
                                     do jj=1,i-1
                                       ns=ns+nr(jj)
                                     enddo
                                endif
                             nss=nr(i)
           do j=1,nss
              if(mlf(j+ns).eq.icl) then
                  ix=ix+1
                    mxyz1(ix)=menot(j+ns)
                    mxyz1(ix+1)=menot(j+ns)
                    mxyz1(ix+2)=menot(j+ns)
                  ix=ix+2
              endif
              if(mlf(j+ns).ne.icl) then
                  iy=iy+1
                    mxyz2(iy)=menot(j+ns)
              endif
           enddo
         enddo
c
c ECRITURE DU NUMERO DE L'OBSERV. ET DES COORD. DES PTS DE SON NUAGE
c
              write(12,*)' '
              write(12,*)' '
              write(12,*)'!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! '
              write(12,5110)icl
              write(12,*)'!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! '
              write(12,*)' '
              write(12,*)'            COORDONNEES '
              write(12,*)' DES POINTS DU NUAGE POUR CE CLIN.'
              write(12,*)' '
         do i=1,ix
              write(12,*)mxyz1(i),mxyz2(i)
         enddo
        enddo
              write(12,*)' '
              write(12,*)' '
c
```

```
c          call duo(nnnn1,nc2,kq)
c
           call part(nbdiad,kq,ngg,ng,nc2)
c
c ECRITURE DU NOMBRE D'EVALUATIONS POSEES PAR CHAQUE OBSERVATEUR
c
          write(12,*)'Nb de diag poses pour 1 etude par chaque clin'
          write(12,*)(nh(i),i=1,nbcl)
          write(12,*)' '
c
c ECRITURE DES DIFFERENTES EVALUATIONS POSEES
c
          write(12,*)'Les diag. diff. poses'
          write(12,*)(ngg(i),i=1,nbdiad)
          write(12,*)' '
c
c ECRITURE DU NOMBRE D'EVALUATIONS DIFFERENTES POSEES
c
          write(12,*)'Nb de diag. diff. poses'
          write(12,*)nbdiad
          write(12,*)' '
          write(12,*)' '
c
c CALCUL DES PARTITIONS POUR L'ENSEMBLE DES SUJETS
c
       nkq=kq/4
       write(12,*)' '
         write(12,*)'                LES n PARTITIONS ',' n=',nkq
         write(12,*)' '
         write(12,*)' '
c                        initialisations:
c
                         par4=0.
                         xva=0.
                         xvaco=0.
                         xdk=0.
                         xdkco=0.
          do i=1,nbpt
                           nss=nr(i)
              do j=1,nss
                 nc1(j)=ndiag(i,j)
              enddo
c
                    call  prime(nc1,nss)
                    call  part(nddif,nss,nyy,nu,nc1)
                    call  duo(nyy,nu,nddif)
c
c ECRITURE DES n PARTITIONS ET DES EVALUATIONS
c ASSOCIES A LEURS DIFFERENTES COMPOSANTES
c
                              write(12,5102)(nyy(j),j=1,nddif)
                              write(12,5103)nxx(i),(nu(j),j=1,nddif)
          if(nss.gt.1) then
c
c
c LES 5 INDICES DE CLASSEMENT DES PARTITIONS CONSIDERES:
c          _ LA VARIANCE _ LA VARIANCE CORRIGEE
c          _ LA CONCENTRATION _ LA CONCENTRATION CORRIGEE
c          _ LE PAR (le coefficient de cartwright)
c
c CALCUL DES INDICES DE CLASSEMENT POUR LES n SUJETS
c
                    call indicev(nddif,nss,nu,xva)
                    call indicevco(nddif,nss,nu,xvaco)
                    call indicdk(nddif,nss,nu,xdk)
                    call indicdkco(nddif,nss,nu,xdkco)
                    call indicepar(nddif,nss,nu,par4)
                              xv1(i)=xva
                              xv2(i)=xvaco
                              xv3(i)=xdk
                              xv4(i)=xdkco
                              xv5(i)=par4
          endif
        enddo
c
c CALCUL DES 5 INDICES DE CLASSEMENT THEORIQUES DES PARTITIONS
c LORSQUE d=4
c
                         z9(1)='1 1 1 1'
                         z9(2)='  1 1 2'
                         z9(3)='    2 2'
                         z9(4)='    1 3'
                         z9(5)='      4'
c
c
        not22(1)=1
        not22(2)=1
        not22(3)=2
                  nddif=3
                  nss=4
                  call indicev(nddif,nss,not22,xva)
                  call indicevco(nddif,nss,not22,xvaco)
                  call indicdk(nddif,nss,not22,xdk)
                  call indicdkco(nddif,nss,not22,xdkco)
                va(1)=0.
                vaco(1)=0.
                dk(1)=2./float(nss**2)
                dkco(1)=0.
                           cartw(1)=0.
                va(2)=xva
                vaco(2)=xvaco
                dk(2)=xdk
                dkco(2)=xdkco
                           cartw(2)=(1./6.)
                va(3)=0.
                vaco(3)=0.
                dk(3)=.5+2./16.+1./16.
                dkco(3)=0.
                           cartw(3)=(1./3.)
                va(4)=(float(nss)/2.-1)**2
                vaco(4)=(float(nss)/2.-1)**2*(1.-(1./3.)**2)
                dk(4)=.5+2.5/16.+1./16.
                dkco(4)=(.5+2.5/16.+1./16.)*(1.-(1./3.)**2)
                           cartw(4)=(1./2.)
                va(5)=(float(nss)/2.-1)**2+1.
                vaco(5)=(float(nss)/2.-1)**2+1.
                dk(5)=1.+1./16.
                dkco(5)=1.+1./16.
                           cartw(5)=1.
              n=5
                                        write(12,*)' '
                                        write(12,*)' '
              call  permu(z9,va,vaco,dk,dkco,cartw,n)
c
```

```
c (les partitions sont ordonnees d'apres leurs variances)
c
      write(12,*)' Partition      Var Var corr',
     .' Concen Concen corr PAR'
      write(12,*)' ---------      --- --------',
     .' ------ ---------- ---'
c
                                    do i=1,5
        write(12,5101)z9(i),va(i),vaco(i),dk(i),dkco(i),cartw(i)
                                    enddo
                                        write(12,*)' '
                                        write(12,*)' '
c
c CALCUL DES PROBABILITES BINOMIALES (n, 1/4) ET (n, 1/2)
c ET DES PROBABILITES BINOMIALES CUMULEES ASSOCIEES
c
       write(12,*)' '
       write(12,*)'  T A B L E S '
       write(12,*)' '
       write(12,*)'
     .                          LOI BINOMIALE (n,1/4)',
       write(12,*)'          LOI BINOMIALE (n,1/2)'
       write(12,*)'        n=',nkq
       write(12,*)'        valeur      prob.        prob. cumulees',
     .                 '
     .            prob.       prob. cumulees'
       write(12,*)' '
c
         n9=kq/4
         x2=0.
         x3=0.
         x5=0.
         x6=0.
         do k9=0,n9-1
            x1=k9+1
            x4=k9+1
           do i=k9+1,n9
                x1=(float(i+1)*x1*3.)/float((n9-i+1)*4)
                x4=(float(i+1)*x4*1.)/float((n9-i+1)*2)
           enddo
                x1=(x1/4.**k9)/float(n9+1)
                x4=(x4/2.**k9)/float(n9+1)
            x2=x2+x1
            x5=x5+x4
c
c ECRITURE DES TABLES DE PROBABILITES  BINOMIALES (n, 1/4) ET (n, 1/2)
c
                          write(12,*)k9,'    ',x1,x2,'   ',x4,x5
                 x9(k9+1)=x2
                 x9(n9+1)=1.
                 x10(k9+1)=x5
                 x10(n9+1)=1.
           x3=x3+x1
           x6=x6+x4
         enddo
           x3=x3+(1./4.)**n9
           x6=x6+(1./2.)**n9
           cnn=(1./4.)**n9
           cn=(1./2.)**n9
                       write(12,*)n9,'    ',cnn,x3,'   ',cn,x6
                                  write(12,*)' '
c                                 write(12,*)' '
```

```
c INDICE DE CLASSEMENT : LA VARIANCE
c --------------------   ===========
            do i=1,nbpt
               xnote(i)=xv1(i)
            enddo
c
c ECRITURE DES VALEURS DE L'INDICE DE CLASSEMENT POUR LES n SUJETS
c
     write(12,*)' '
     write(12,*)'-------- '
     write(12,*)' '
     write(12,*)'Aux sujets sont attribuees les variances :'
     write(12,*)' '
                                  write(12,*)(xv1(i),i=1,nbpt)
                                        write(12,*)' '
c
c CALCUL DES FREQUENCES DE PARTITIONS TYPES SUR L'ENSEMBLE DES SUJETS
c
              call  primx(xnote,nbpt)
               ld=0
               nt12=0
               lamd=1
            do i=2,nbpt
              if(xnote(i-1).eq.xnote(i)) then
               lamd=lamd+1
              else
               ld=ld+1
               xnotee(ld)=xnote(i-1)
               not11(ld)=lamd
               nt12=nt12+lamd
               lamd=1
              endif
            enddo
               xnotee(ld+1)=xnote(nbpt)
               not11(ld+1)=nbpt-nt12
c
c ECRITURE DES VALEURS TYPES DE LA VAR. ET LEUR FREQUENCE (non nulle)
c
     write(12,*)' '
     write(12,*)' '
     write(12,*)'Les variances types et leur frequence :'
     write(12,*)' '
          write(12,5104)(xnotee(i),i=1,ld+1)
                                        write(12,*)' '
          write(12,5105)(not11(i),i=1,ld+1)
                                        write(12,*)' '
c
c TESTS SUR LA STATISTIQUE FMAX; LES RESULTATS
c
     do i=1,ld+1
        mu0=not11(i)
        bor2=float(nbpt)*(1.-x9(mu0+1))
        bor1=1.-exp(float(-nbpt)*(1.-x9(mu0+1)))
      write(12,5106)mu0
      write(12,5107)bor1,bor2
     enddo
                                        write(12,*)' '
c
c TESTS SUR LA STATISTIQUE BINOMIALE (n, 1/2); LES RESULTATS
c
       mu1=not11(ld+1)
       mu2=not11(ld)+not11(ld+1)
```

```
      xx101=1.-x10(mu1+1)
      xx102=1.-x10(mu2+1)
      write(12,5108)mu1,xx101
      write(12,5109)
      write(12,5108)mu2,xx102
      write(12,5109)
                                        write(12,*)' '
                                        write(12,*)' '
c
c INDICE DE CLASSEMENT : LA VARIANCE CORRIGEE
c --------------------   ====================
            do i=1,nbpt
               xnote(i)=xv2(i)
            enddo
c
c ECRITURE DES VALEURS DE L'INDICE DE CLASSEMENT POUR LES n SUJETS
c
      write(12,*)' '
      write(12,*)'-------- '
      write(12,*)' '
      write(12,*)'Aux sujets sont attribuees les variances corrigees :'
      write(12,*)' '
                                  write(12,*)(xv2(i),i=1,nbpt)
                                        write(12,*)' '
c
c CALCUL DES FREQUENCES DE PARTITIONS TYPES SUR L'ENSEMBLE DES SUJETS
c
             call  primx(xnote,nbpt)
              ld=0
              nt12=0
              lamd=1
            do i=2,nbpt
              if(xnote(i-1).eq.xnote(i)) then
               lamd=lamd+1
              else
               ld=ld+1
               xnotee(ld)=xnote(i-1)
               not12(ld)=lamd
               nt12=nt12+lamd
               lamd=1
              endif
            enddo
               xnotee(ld+1)=xnote(nbpt)
               not12(ld+1)=nbpt-nt12
c
c ECRITURE DES VALEURS TYPES DE LA VAR. CORR. ET LEUR FREQ. (non nulle)
c
      write(12,*)' '
      write(12,*)' '
      write(12,*)'Les variances corrigees types et leur frequence :'
      write(12,*)' '
           write(12,5104)(xnotee(i),i=1,ld+1)
                                        write(12,*)' '
           write(12,5105)(not12(i),i=1,ld+1)
                                        write(12,*)' '
c
c TESTS SUR LA STATISTIQUE FMAX; LES RESULTATS
c
      do i=1,ld+1
         mu0=not12(i)
         bor2=float(nbpt)*(1.-x9(mu0+1))
         bor1=1.-exp(float(-nbpt)*(1.-x9(mu0+1)))
```

```
      write(12,5106)mu0
      write(12,5107)bor1,bor2
     enddo
                                        write(12,*)' '
c
c TESTS SUR LA STATISTIQUE BINOMIALE (n, 1/2); LES RESULTATS
c
      mu1=not12(ld+1)
      mu2=not12(ld)+not12(ld+1)
      xx101=1.-x10(mu1+1)
      xx102=1.-x10(mu2+1)
      write(12,5108)mu1,xx101
      write(12,5109)
      write(12,5108)mu2,xx102
      write(12,5109)
                                        write(12,*)' '
                                        write(12,*)' '
c
c INDICE DE CLASSEMENT : LA CONCENTRATION
c --------------------   ================
            do i=1,nbpt
               xnote(i)=xv3(i)
            enddo
c
c ECRITURE DES VALEURS DE L'INDICE DE CLASSEMENT POUR LES n SUJETS
c
      write(12,*)' '
      write(12,*)'-------- '
      write(12,*)' '
      write(12,*)'Aux sujets sont attribuees les concentrations :'
      write(12,*)' '
                                  write(12,*)(xv3(i),i=1,nbpt)
                                        write(12,*)' '
c
c CALCUL DES FREQUENCES DE PARTITIONS TYPES SUR L'ENSEMBLE DES SUJETS
c
             call  primx(xnote,nbpt)
              ld=0
              nt12=0
              lamd=1
            do i=2,nbpt
              if(xnote(i-1).eq.xnote(i)) then
               lamd=lamd+1
              else
               ld=ld+1
               xnotee(ld)=xnote(i-1)
               not13(ld)=lamd
               nt12=nt12+lamd
               lamd=1
              endif
            enddo
               xnotee(ld+1)=xnote(nbpt)
               not13(ld+1)=nbpt-nt12
c
c ECRITURE DES VALEURS TYPES DE LA CONCENTRATION
c ET LEUR FREQUENCE (non nulle)
c
      write(12,*)' '
      write(12,*)' '
      write(12,*)'Les concentrations types et leur frequence :'
      write(12,*)' '
           write(12,5104)(xnotee(i),i=1,ld+1)
```

```
                                        write(12,*)' '
      write(12,5105)(not13(i),i=1,ld+1)
                                        write(12,*)' '
c
c TESTS SUR LA STATISTIQUE FMAX; LES RESULTATS
c
      do i=1,ld+1
         mu0=not13(i)
         bor2=float(nbpt)*(1.-x9(mu0+1))
         bor1=1.-exp(float(-nbpt)*(1.-x9(mu0+1)))
       write(12,5106)mu0
       write(12,5107)bor1,bor2
      enddo
                                        write(12,*)' '
c
c TESTS SUR LA STATISTIQUE BINOMIALE (n, 1/2); LES RESULTATS
c
       mu1=not13(ld+1)
       mu2=not13(ld)+not13(ld+1)
       xx101=1.-x10(mu1+1)
       xx102=1.-x10(mu2+1)
       write(12,5108)mu1,xx101
       write(12,5109)
       write(12,5108)mu2,xx102
       write(12,5109)

                                        write(12,*)' '
                                        write(12,*)' '
c
c INDICE DE CLASSEMENT : LA CONCENTRATION CORRIGEE
c --------------------   ==========================
              do i=1,nbpt
                 xnote(i)=xv4(i)
              enddo
c
c ECRITURE DES VALEURS DE L'INDICE DE CLASSEMENT POUR LES n SUJETS
c
      write(12,*)' '
      write(12,*)'-------- '
      write(12,*)' '
      write(12,*)'Aux sujets sont attribuees les concentrations corr. :'
      write(12,*)' '
                                  write(12,*)(xv4(i),i=1,nbpt)
                                  write(12,*)' '
c
c CALCUL DES FREQUENCES DE PARTITIONS TYPES SUR L'ENSEMBLE DES SUJETS
c
               call  primx(xnote,nbpt)
                ld=0
                nt12=0
                lamd=1
              do i=2,nbpt
                if(xnote(i-1).eq.xnote(i)) then
                 lamd=lamd+1
                else
                 ld=ld+1
                 xnotee(ld)=xnote(i-1)
                         xdb(ld)=xnote(i-1)
                 not14(ld)=lamd
                 nt12=nt12+lamd
                 lamd=1
                endif
              enddo
```

```
              xnotee(ld+1)=xnote(nbpt)
                      xdb(ld+1)=xnote(nbpt)
              not14(ld+1)=nbpt-nt12
c
c ECRITURE DES VALEURS TYPES DE LA CONCENTRATION CORRIGEE
c ET LEUR FREQUENCE (non nulle)
c
     write(12,*)' '
     write(12,*)' '
     write(12,*)'Les concentrations corr. types et leur frequence :'
     write(12,*)' '
          write(12,5104)(xnotee(i),i=1,ld+1)
                                        write(12,*)' '
          write(12,5105)(not14(i),i=1,ld+1)
                                        write(12,*)' '
c
c TESTS SUR LA STATISTIQUE FMAX; LES RESULTATS
c
     do i=1,ld+1
        mu0=not14(i)
        bor2=float(nbpt)*(1.-x9(mu0+1))
        bor1=1.-exp(float(-nbpt)*(1.-x9(mu0+1)))
      write(12,5106)mu0
      write(12,5107)bor1,bor2
     enddo
                                        write(12,*)' '
c
c TESTS SUR LA STATISTIQUE BINOMIALE (n, 1/2); LES RESULTATS
c
      mu1=not14(ld+1)
      mu2=not14(ld)+not14(ld+1)
      xx101=1.-x10(mu1+1)
      xx102=1.-x10(mu2+1)
      write(12,5108)mu1,xx101
      write(12,5109)
      write(12,5108)mu2,xx102
      write(12,5109)
                                        write(12,*)' '
                                        write(12,*)' '
c
c INDICE DE CLASSEMENT : LE PAR
c --------------------   ======
             do i=1,nbpt
                xnote(i)=xv5(i)
             enddo
c
c ECRITURE DES VALEURS DE L'INDICE DE CLASSEMENT POUR LES n SUJETS
c
     write(12,*)' '
     write(12,*)'-------- '
     write(12,*)' '
     write(12,*)'Aux sujets sont attribuees les PAR :'
     write(12,*)' '
                                  write(12,*)(xv5(i),i=1,nbpt)
                                  write(12,*)' '
c
c CALCUL DES FREQUENCES DE PARTITIONS TYPES SUR L'ENSEMBLE DES SUJETS
c
               call  primx(xnote,nbpt)
                ld=0
                nt12=0
                lamd=1
```

```
                do i=2,nbpt
                  if(xnote(i-1).eq.xnote(i)) then
                   lamd=lamd+1
                  else
                   ld=ld+1
                   xnotee(ld)=xnote(i-1)
                   not15(ld)=lamd
                   nt12=nt12+lamd
                   lamd=1
                  endif
                enddo
                   xnotee(ld+1)=xnote(nbpt)
                   not15(ld+1)=nbpt-nt12
c
c ECRITURE DES VALEURS TYPES DU PAR ET LEUR FREQUENCE (non nulle)
c
      write(12,*)' '
      write(12,*)' '
      write(12,*)'Les PAR types et leur frequence :'
      write(12,*)' '
           write(12,5104)(xnotee(i),i=1,ld+1)
                                        write(12,*)' '
           write(12,5105)(not15(i),i=1,ld+1)
                                        write(12,*)' '
                                        write(12,*)' '
c
c CALCULS DES COEFFICIENTS D'ACCORD
c
           accord1=0.
           accord2=0.
           accord3=0.
           accord4=0.
           accord5=0.
           accord6=0.
         do i=1,4
           xniveau(i)=float(i-1)/3.
            accord1=accord1+xniveau(i)*float(not11(i))
            accord2=accord2+xniveau(i)*float(not12(i))
            accord3=accord3+xniveau(i)*float(not13(i))
            accord4=accord4+xniveau(i)*float(not14(i))
            accord5=accord5+xnotee(i)*float(not15(i))
            accord6=accord6+xdb(i)*float(not14(i))
         enddo
           accord1=accord1/float(nbpt)
           accord2=accord2/float(nbpt)
           accord3=accord3/float(nbpt)
           accord4=accord4/float(nbpt)
           accord5=accord5/float(nbpt)
           accord6=accord6/float(nbpt)
        write(12,*)' '
        write(12,*)' '
     write(12,*)'Indice d accord pour la variance :          ',accord1
        write(12,*)' '
     write(12,*)'Indice d accord pour la variance corr. :    ',accord2
        write(12,*)' '
     write(12,*)'Indice d accord pour la concentration :     ',accord3
        write(12,*)' '
     write(12,*)'Indice d accord pour la concentration corr. :',accord4
        write(12,*)' '
     write(12,*)'Indice d accord pour le PAR :               ',accord5
        write(12,*)' '
        write(12,*)' '
     write(12,*)'Coefficient de concentration moyen :        ',accord6
        write(12,*)' '
        write(12,*)' '
        write(12,*)' '
        write(12,*)'                                    * * * '
        write(12,*)' '
c
c
5101       format(2x,a7,' : ',2x,6(f6.3,2x))
5102    format(1x,16x,'Diagnostics:',20('----',i1))
5103    format(1x,' Patient no:',i4,'; Partition:',20(4x,i1))
5104    format(4x,20f6.3)
5105    format(4x,20(i3,3x))
5106    format(2x,'La prob. pour Fmax d exceder la frequence:',i4)
5107    format(2x,'       est comprise entre les deux bornes:',
     .          4x,f7.3,'  et ',f7.3)
5108    format(2x,'F  excede ',i3,'   avec probabilite  'f6.3)
5109    format(3x,'I')
5110    format(9x,'Clinicien no:',i4)
 200    format(1x/1x,'  Nb de cliniciens'
     .          ,'  Nb de patients','  Taille des donnees')
2000    format(10x,i3,13x,i4,13x,i4)
c
         stop
         end
c
c
      subroutine  prime(nxt,ny)
         integer nxt(ny)
             if(ny.gt.1)  then
               do  i1=1,ny-1
                    do  i2=i1,ny
                  if(nxt(i1).gt.nxt(i2))  then
                      cx=nxt(i2)
                      nxt(i2)=nxt(i1)
                      nxt(i1)=cx
                  endif
                    enddo
               enddo
             endif
                         return
                         end
c
c QUATRE SUBROUTINES DE CLASSEMENT
c
      subroutine  primx(xnxt,ny)
         real xnxt(ny)
             if(ny.gt.1)  then
               do  i1=1,ny-1
                    do  i2=i1,ny
                  if(xnxt(i1).gt.xnxt(i2))  then
                      cx=xnxt(i2)
                      xnxt(i2)=xnxt(i1)
                      xnxt(i1)=cx
                  endif
                    enddo
               enddo
             endif
                         return
                         end
c
```

```
      subroutine  duo(kx,nxt,ny)
         integer nxt(ny),kx(ny)
             if(ny.gt.1)  then
               do  i1=1,ny-1
                    do  i2=i1,ny
                  if(nxt(i1).gt.nxt(i2))  then
                      cx=nxt(i2)
                      nxt(i2)=nxt(i1)
                      nxt(i1)=cx
                        cx=kx(i2)
                        kx(i2)=kx(i1)
                        kx(i1)=cx
                  endif
                    enddo
               enddo
             endif
                          return
                          end
c
      subroutine  tert(kx,nxt,mx,ny)
         integer nxt(ny),kx(ny),mx(ny)
             if(ny.gt.1)  then
               do  i1=1,ny-1
                    do  i2=i1,ny
                  if(nxt(i1).gt.nxt(i2))  then
                      cx=nxt(i2)
                      nxt(i2)=nxt(i1)
                      nxt(i1)=cx
                        cx=kx(i2)
                        kx(i2)=kx(i1)
                        kx(i1)=cx
                      cx=mx(i2)
                      mx(i2)=mx(i1)
                      mx(i1)=cx
                  endif
                    enddo
               enddo
             endif
                          return
                          end
c
          subroutine  permu(w,nxt,kx,mx,nxch,nxper,ny)
              character*100 w(ny),ccx
       real nxt(ny),kx(ny),mx(ny),nxch(ny),cx,nxper(ny)
           if(ny.gt.1)  then
             do  i1=1,ny-1
                  do  i2=i1,ny
                if(nxt(i1).gt.nxt(i2))  then
                      ccx=w(i2)
                      w(i2)=w(i1)
                      w(i1)=ccx
                   cx=nxt(i2)
                   nxt(i2)=nxt(i1)
                   nxt(i1)=cx
                     cx=kx(i2)
                     kx(i2)=kx(i1)
                     kx(i1)=cx
                   cx=mx(i2)
                   mx(i2)=mx(i1)
                   mx(i1)=cx
                     cx=nxch(i2)
                     nxch(i2)=nxch(i1)
                     nxch(i1)=cx
                   cx=nxper(i2)
                   nxper(i2)=nxper(i1)
                   nxper(i1)=cx
c                     cx=mz(i2)
c                     mz(i2)=mz(i1)
c                     mz(i1)=cx
                  endif
                    enddo
               enddo
             endif
                          return
                          end
c
c SUBROUTINE DE CALCUL DES COMPOSANTES D'UNE PARTITION
c
          subroutine  part(ka,kq,nxx,nr,nx)
                        integer nxx(ka),nr(ka),nx(kq)
                 nxx(1)=nx(1)
                  nr(1)=1
                     ka=1
                     kb=1
             if(kq.gt.1) then
                 do i=1,kq-1
                     if(nx(i).eq.nx(i+1)) then
                        kb=kb+1
                        nr(ka)=kb
                     else
                        ka=ka+1
                        nr(ka)=1
                        kb=1
                        nxx(ka)=nx(i+1)
                     endif
                 enddo
             endif
                          return
                          end
c
c LES SUBROUTINES DE CALCUL DES CINQ INDICES DE CLASSEMENT DES PARTIONS
c
          subroutine  indicepar(nddif,nss,nu,par)
                        integer nu(nddif)
                          annu=0
                          annumax=1
                   if(nss.ge.2) then
                      do ir=1,nddif
                        if(nu(ir).ge.2) then
                          annu=annu+(float(nu(ir)*(nu(ir)-1)))/2.
                        endif
                      enddo
                    endif
                    if(nss.ge.2)annumax=(float(nss*(nss-1)))/2.
                         par=annu/annumax
          return
          end
c
          subroutine  indicev(nddif,nss,nu,var)
                        integer nu(nddif)
               vv=((float(nss))/2.-1.)**2
                     espnu=0.
                     varnu=0.
                  do i=1,nddif
```

```
                espnu=espnu+float(nu(i))
              enddo
c
c
             espnu=espnu/(float(nddif))

                do i=1,nddif
                   varnu=varnu+((float(nu(i))-espnu))**2
                enddo
c
             varnu=varnu/(float(nddif))
          var=varnu
                if(nddif.eq.1) then
                          var=(float(nss)/2.-1.)**2+1.
                endif
c
          xcorr=1.
c
             var=var*xcorr
          return
          end
c
          subroutine  indicevco(nddif,nss,nu,var)
                      integer nu(nddif)
               vv=((float(nss))/2.-1.)**2
                     espnu=0.
                     varnu=0.
                  do i=1,nddif
                     espnu=espnu+float(nu(i))
                  enddo
c
               espnu=espnu/(float(nddif))
c
                  do i=1,nddif
                     varnu=varnu+((float(nu(i))-espnu))**2
                  enddo
c
               varnu=varnu/(float(nddif))
          var=varnu
                if(nddif.eq.1) then
                          var=(float(nss)/2.-1.)**2+1.
                endif
c
             if(nddif.eq.1)xcorr=1.
             if(nddif.gt.1) then
              xcorr=1.-((float(nu(nddif-1))/float(nu(nddif)))**2)
             endif
c
             var=var*xcorr
          return
          end
c
          subroutine  indicdk(nddif,nss,nu,dk)
                      integer nu(nddif)
             so=0.
           do j=1,nddif
             so=so+nu(j)**2
           enddo
             dk=-(float(nddif)/float(nss))+(1./float(nss)**2)
                +(1./float(nss)**3)*so+1.
          return
          end
c
```

```
          subroutine  indicdkco(nddif,nss,nu,dk)
                      integer nu(nddif)
             so=0.
           do j=1,nddif
             so=so+nu(j)**2
           enddo
             dk=-(float(nddif)/float(nss))+(1./float(nss)**2)
                +(1./float(nss)**3)*so+1.
             if(nddif.eq.1)xcorr=1.
             if(nddif.gt.1) then
              xcorr=1.-((float(nu(nddif-1))/float(nu(nddif)))**2)
             endif
          dk=dk*xcorr
          return
          end
```

PROGRAMME DE CALCULS DE CONCENTRATIONS EN DIM 2

RESUME DU PROGRAMME : CONCENT.FOR

```
c """""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""
c
c ENTREE DE LA MATRICE DE DONNEES n*2, EN FORMAT LIBRE, DANS UN
c FICHIER
c
c
c ECRITURE DE LA MATRICE DE DONNEES X'
c
c
c ENTREES EN INTERACTIVITE:
c de M, nombre de termes de la serie tronquee,
c de n, nombre de points
c de N, la puissance de la precision dans l'objectif
c
c
c NORMES DES COLONNES DE LA MATRICE DE DONNEES
c
c
c CLASSEMENT DES n NORMES DE LA PLUS PETITE A LA PLUS GRANDE
c
c
c BOUCLE SUR LES DIFFERENTES VALEURS DE LA FONCTION OBJECTIF
c
c
c      PENTE DE LA PROJECTION
c
c      CALCUL DES POINTS PROJETES
c
c      LES TERMES DE LA SERIE INFINIE DU COEFF DE CONCENT
c
c
c FIN DE LA BOUCLE
c
c
c ECRITURES DE LA PENTE ET DE LA FONCTION OBJECTIF
c
c
c LES CONCENTRATIONS MAXIMUM ET MINIMUM
c
c
c POINT PAR OU PASSE LA DROITE DE PLUS FORTE CONCENTRATION
c
c
c POINT PAR OU PASSE LA DROITE ORTHOGONALE A LA DROITE DE PLUS
c FORTE CONCENTRATION
c
c
c POINT PAR OU PASSE LA DROITE DE PLUS FAIBLE CONCENTRATION
c
c
c SORTIE DES RESULTATS DANS UN FICHIER
c
c
      stop
      end
c
c
c DEUX SUBROUTINES DE CLASSEMENT
c
c """""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""""
```

PROGRAMME CONCENT.FOR , ECRIT EN FORTRAN 77

```
      real*8 xx(2,100),z(100),w(100),x(500),pente(500),objfun(500)
     .      ,xxnorm(100)
     .  ,g,g1,g2,g3,g11,g22,g2tot,g3tot,y1,y2,y3,xk,xk1,xk2,cx,xmax
     .  ,sr,tr,ts,tt,tx,ty,pent
      real a1,a2,a3,a4,a5,a6
      integer*4 mm,nn,nnn,nh
      integer n1(500),n2(500)
c
      open(file='re.fun',unit=8,status='new')
      open(file='do.fun',unit=12,status='old')
c
c ENTREE DE LA MATRICE DE DONNEES n*2, EN FORMAT LIBRE, DANS LE FICHIER
c
c                        ----- do.fun -----
c
c SORTIE DES RESULTATS DANS LE FICHIER
c
c                        ----- re.fun -----
c
c
c
c ENTREES
c
c    1) La matrice de donnees n*2, en format libre, dans le fichier: do.
fun
c    2) Nb de termes de la serie tronquee, nb de points, precision (N)
c       en interactivite
c
c SORTIES
c
c    1) La matrice de donnees
c    2) 101 couples: pente de projection-valeur de l'objectif
c    3) Les concentrations maximum et minimum
c    4) La pente et un point de la droite de concentration max.
c    5) La pente et un point de la droite orthog a la droite de concent
max.
c    6) La pente et un point de la droite de concentration min.
c
c
c
c ENTREES EN INTERACTIVITE
c
      print*,' '
      print*,'M, nb de termes de   n, nb de points; N, puissance de la'
     .   ,' precision'
      print*,'  la serie tronquee '
      print*,' '
      print*,'Donner M, n, et N '
      print*,'(Suggestions :_1)  M=3 si n<30, sinon M=2  _2)  N=10) '
      print*,' '
      read(5,*)mm,nn,nnn
      print*,' '
            mm=mm-1
c
       write(8,*)' '
       write(8,*)' '
       write(8,*)'_____________________________________________________'
       write(8,*)'|_____________________________________________________|'
       write(8,*)'| PROGRAMME DE CALCULS DE CONCENTRATIONS EN DIM 2 |'
       write(8,*)'|_____________________________________________________|'
       write(8,*)' '
       write(8,*)' '
       write(8,*)' '
       write(8,*)' '
c
c
c ECRITURE DE M, n, N
c
c
       write(8,*)' '
       write(8,*)' '
       write(8,*)' '
                                                          mmfin=mm+1
       write(8,*)'    Nb de termes de la serie M :  ',mmfin
       write(8,*)' '
                                                          nnfin=nn
       write(8,*)'    Nb de points dans le nuage n :',nnfin
       write(8,*)' '
                                                          nnnfin=nnn
       write(8,*)'    Puissance de la precision N : ',nnnfin
       write(8,*)' '
       write(8,*)' '
       write(8,*)' '
c
c
       write(8,*)'           LA MATRICE DE DONNEES N*2'
       write(8,*)' '
c
c ENTREE DE LA MATRICE DE DONNEES n*2
c
      do j=1,nn
       read(12,*)xx(1,j),xx(2,j)
      enddo
c
c
c ECRITURE DE LA MATRICE DE DONNEES X' (n*2):
c
       do j=1,nn
        write(8,*)xx(1,j),xx(2,j)
       enddo
        write(8,*)' '
        write(8,*)' '
        write(8,*)' '
      write(8,*)' LA FONCTION OJECTIF CALCULEE EN 101 POINTS'
      write(8,*)' POUR 101 PENTES DE LA DROITE DE PROJECTION'
        write(8,*)' '
      write(8,*)'La pente de projection  La fonction objectif '
      write(8,*)'========================  ====================== '
        write(8,*)' '
c
c NORMES DES COLONNES DE LA MATRICE DE DONNEES
c
      do j=1,nn
         xxnorm(j)=sqrt(xx(1,j)**2+xx(2,j)**2)
      enddo
c
c
c CLASSEMENT DES n NORMES DE LA PLUS PETITE A LA PLUS GRANDE
c
                    call  prime(xxnorm,nn)
c
```

```
c
c BOUCLE SUR LES DIFFERENTS POINTS DE LA FONCTION
c
      do nx=1,101
       x(nx)=float(nx-1)/50.-1.
c
       sr=sqrt(1.-x(nx)**2)
c
c PENTE DE LA PROJECTION
c
      if(x(nx).ne.0.) pente(nx)=sr/x(nx)
c
c                                          Initialisation :
                                           objfun(nx)=.0
c
c LES TERMES DE LA SERIE INFINIE DU COEFF DE CONCENT
c
      do m=0,mm
c                                          Initialisation :
                                           g=.0
c
         do nu=1,nn**m
c**********************************1er terme de la fnct objectif
c                                          Initialisation :
                                           g11=0.
            do ni=1,nn
c
c Integration de la contrainte dans la fonction : y**2=1-x**2
c On decide de prendre toujours y positif (sqrt donne une racine >0)
c UN POINT PROJETE:
c
                    z(ni)=x(nx)*xx(1,ni)+sr*xx(2,ni)
c
c       tr=nn**(m+2)
c         w(ni)=(z(ni)*tr)/(float(2*(nn+1))*xmax)
c     .                +tr/float(2*(nn+1))-float(nn*(nu-1))
       tt=nn**(m+1)
         w(ni)=(z(ni)*tt)/(2.*xmax+.00000001)
     .               +(tt*xmax)/(2.*xmax+.00000001)-float(nn*(nu-1))
       ts=nn
       tx=(2.*w(ni)-ts)/ts
             if((tx.lt.1.).and.(tx.ge.-1.)) ty=.00000001
             if((tx.ge.1.).or.(tx.lt.-1.)) ty=10.
       g11=g11+1./(1.+(w(ni)/(w(ni)-ts))**2
     .                *ty**(nnn*2))
             enddo
      g1=-(g11**2)
c**********************************2eme terme de la fnct objectif
c                         Initialisation :
                          g2tot=0.
             do k=0,nn-1
c                         Initialisation :
                          g2=0.
                do ni=1,nn
          xk=k
          xk1=k+1
          xk2=2*k+1
      tx=2.*w(ni)-xk2
             if((tx.lt.1.).and.(tx.ge.-1.)) y3=.00000001
             if((tx.ge.1.).or.(tx.lt.-1.)) y3=10.
          y1=(w(ni)-xk1)**2
          y2=(w(ni)-xk)**2
                          g2=g2+y1/(y1+y2*y3**(nnn*2))
                enddo
               g2tot=g2tot+g2
             enddo
      g2=g2tot*g11
c**********************************3eme terme de la fnct objectif
c                         Initialisation :
                          g3=0.
             do k=0,nn-1
c             Initialisation :
                          g22=0.
                do ni=1,nn
          xk=k
          xk1=k+1
          xk2=2*k+1
      tx=2.*w(ni)-xk2
             if((tx.lt.1.).and.(tx.ge.-1.)) y3=.00000001
             if((tx.ge.1.).or.(tx.lt.-1.)) y3=10.
          y1=(w(ni)-xk1)**2
          y2=(w(ni)-xk)**2
                          g22=g22+y1/(y1+y2*y3**(nnn*2))
                enddo
                          g3tot=1.
               do nh=0,nn-1
             g3tot=(g3tot*(float(nn-nh)-g22))/float(nn-nh)
               enddo
             g3=g3+g3tot
            enddo
         g3=g11*g3
         g=g+g1+g2+g3
c
          enddo
cc enddo pour la variation de u
c
         objfun(nx)=objfun(nx)+g/float(nn**(m+2))
c
      enddo
cc enddo pour la variation de m
c
      enddo
cc enddo pour la variation des projections a l'aide d'x(nx)
c
c
c
c ECRITURES DE LA PENTE ET DE LA FONCTION OBJECTIF
c
      do nx=1,101
       if(x(nx).ne.0.) write(8,*)pente(nx),objfun(nx)
       if(x(nx).eq.0.) write(8,9001)objfun(nx)
      enddo
c
c LES CONCENTRATIONS MAXIMUM ET MINIMUM
c
            nb=101
          call  permu(pente,x,objfun,nb)
      write(8,*)' '
      write(8,*)' '
      write(8,*)'La concentration max. ; La concentration min.'
      write(8,*)' '
      print*,' '
      print*,' '
```

```
      print*,'La concentration max. ; La concentration min.'
      print*,' '
       write(8,*)objfun(nb),objfun(1)
      write(8,*)'*********************   ********************* '
      write(8,*)' '
      write(8,*)' '
       print*,objfun(nb),objfun(1)
      print*,'*********************   ********************* '
      print*,' '
      print*,' '
c
c POINT PAR OU PASSE LA DROITE DE PLUS FORTE CONCENTRATION
c
          do ni=1,nn
            z(ni)=sqrt(1.-x(nb)**2)*xx(1,ni)-x(nb)*xx(2,ni)
            w(ni)=ni
          enddo
c
                call  permu(xxnorm,w,z,nn)
c
            if((float(nn)/2.-nn/2).eq.0.) then
                  nii=nn/2
                  niii=(nn+2)/2
                  kii=w(nii)
                  kiii=w(niii)
              a3=(xx(1,kii)+xx(1,kiii))/2.
              a4=(xx(2,kii)+xx(2,kiii))/2.
            else
                  nii=(nn+1)/2
                  kii=w(nii)
              a3=xx(1,kii)
              a4=xx(2,kii)
            endif
       write(8,*)' '
       write(8,*)'-----------------------------------------------------'
       write(8,*)'La droite de concent. maximum: la pente et un point'
       write(8,*)'-----------------------------------------------------'
       print*,' '
       print*,'-----------------------------------------------------'
       print*,'La droite de concent. maximum: la pente et un point'
       print*,'-----------------------------------------------------'
      if(x(nb).ne.0.)write(8,*)pente(nb),a3,a4
      if(x(nb).eq.0.)write(8,9001)a3,a4
      if(x(nb).ne.0.)print*,pente(nb),a3,a4
      if(x(nb).eq.0.)print*,' pente verticale       ',a3,a4
       write(8,*)' '
       print*,' '
c
c
c POINT PAR OU PASSE LA DROITE ORTHOGONALE A LA DROITE DE PLUS
c FORTE CONCENTRATION
c
          do ni=1,nn
            z(ni)=x(nb)*xx(1,ni)+sqrt(1.-x(nb)**2)*xx(2,ni)
            w(ni)=ni
          enddo
c
                call  permu(xxnorm,w,z,nn)
c
            if((float(nn)/2.-nn/2).eq.0.) then
                  nii=nn/2
                  niii=(nn+2)/2
                  kii=w(nii)
                  kiii=w(niii)
              a1=(xx(1,kii)+xx(1,kiii))/2.
              a2=(xx(2,kii)+xx(2,kiii))/2.
            else
                  nii=(nn+1)/2
                  kii=w(nii)
              a1=xx(1,kii)
              a2=xx(2,kii)
            endif
      if(sqrt(1.-x(nb)**2).ne.0.) pent=-x(nb)/sqrt(1.-x(nb)**2)
       write(8,*)' '
       write(8,*)'-----------------------------------------------------'
       write(8,*)'La droite orthogonale a la droite de plus '
       write(8,*)'       forte concentration : la pente et un point'
       write(8,*)'-----------------------------------------------------'
       print*,' '
       print*,'-----------------------------------------------------'
       print*,'La droite orthogonale a la droite de plus '
       print*,'       forte concentration : la pente et un point'
       print*,'-----------------------------------------------------'
      if(sqrt(1.-x(nb)**2).ne.0.)write(8,*)pent,a1,a2
      if(sqrt(1.-x(nb)**2).eq.0.)write(8,9001)a1,a2
      if(sqrt(1.-x(nb)**2).ne.0.) print*,pent,a1,a2
      if(sqrt(1.-x(nb)**2).eq.0.) print*,' pente verticale       ',a1,a2

       write(8,*)' '
       print*,' '
c
c POINT PAR OU PASSE LA DROITE DE PLUS FAIBLE CONCENTRATION
c
          do ni=1,nn
            z(ni)=sqrt(1.-x(1)**2)*xx(1,ni)-x(1)*xx(2,ni)
            w(ni)=ni
          enddo
c
                call  permu(xxnorm,w,z,nn)
c
            if((float(nn)/2.-nn/2).eq.0.) then
                  nii=nn/2
                  niii=(nn+2)/2
                  kii=w(nii)
                  kiii=w(niii)
              a5=(xx(1,kii)+xx(1,kiii))/2.
              a6=(xx(2,kii)+xx(2,kiii))/2.
            else
                  nii=(nn+1)/2
                  kii=w(nii)
              a5=xx(1,kii)
              a6=xx(2,kii)
            endif
       write(8,*)' '
       write(8,*)'-----------------------------------------------------'
       write(8,*)'La droite de concent. minimum: la pente et un point'
       write(8,*)'-----------------------------------------------------'
       print*,' '
       print*,'-----------------------------------------------------'
       print*,'La droite de concent. minimum: la pente et un point'
       print*,'-----------------------------------------------------'
      if(x(1).ne.0.)write(8,*)pente(1),a5,a6
      if(x(1).eq.0.)write(8,9001)a5,a6
      if(x(1).ne.0.)print*,pente(1),a5,a6
```

```
      if(x(1).eq.0.)print*,' pente verticale         ',a5,a6
       write(8,*)' '
       write(8,*)' '
       write(8,*)'                          *  *  * '
       write(8,*)' '
       print*,' '
       print*,' '
c
c
9001   format(1x,' pente verticale          ',f9.6,2x,f9.6)
c
c
       stop
       end
c
c DEUX SUBROUTINES DE PERMUTATION
c
        subroutine  prime(x,nn)
             real*8 x(nn),cx
               do  i1=1,nn-1
                    do  i2=i1,nn
                       if(x(i1).gt.x(i2))  then
                         cx=x(i2)
                         x(i2)=x(i1)
                         x(i1)=cx
                       endif
               enddo
                    enddo
        return
           end
c
        subroutine  permu(pente,x,objfun,nn)
             real*8 pente(nn),x(nn),objfun(nn),cx
               do  i1=1,nn-1
                    do  i2=i1,nn
                       if(objfun(i1).gt.objfun(i2))  then
                         cx=objfun(i2)
                         objfun(i2)=objfun(i1)
                         objfun(i1)=cx
                          cx=pente(i2)
                          pente(i2)=pente(i1)
                          pente(i1)=cx
                         cx=x(i2)
                         x(i2)=x(i1)
                         x(i1)=cx
                       endif
               enddo
                    enddo
        return
           end
```


QUELQUES RESULTATS EMPIRIQUES ET COMMENTAIRES

(Les sorties numeriques proviennent des programmes Anaconda.for et Concent.for)

Une analyse de concordance extraite d' une etude realisee a l'OMS

On presente,ici, une etude realisee dans le cadre d'un programme de recherche lance par l'OMS (Geneve), utilisant des methodes d'analyse de concordance.

Cette etude a porte sur des donnees provenant de 42 pays.

Pour chaque pays, plusieurs centres ont ete analyses (centres hospitaliers de plusieurs villes).

Le nombre total de centres est de 105; ils sont codes entre les chiffres 100 et les chiffres 700.

Ci-dessous sont presentes les resultats concernant le centre 487:

Centre de therapie breve,Chene-Bourg,Geneve,Suisse,
Service du docteur N. AAPRO

Les evaluations sont communes a tous les centres:

Codees de 0 a 9

Les cliniciens (observateurs) evaluent les patients (sujets) au vu de l'echelle qualitative nominale dont les modalites sont les codes ci-dessus.

num. du centre	num.du clin.	age	sexe
487	1	36	2
487	2	43	2
487	3	44	1
487	4	34	1

Num du centre	Nb de cliniciens	Nb de patients	Taille des donnees
487	4	25	100

L etude de concordance ne prend en compte que :

Nb de cliniciens	Nb de patients
4	25

1ere serie: les num. de patients; 2eme serie: le nb de clin. par pat.

101	102	103	104	105	121
122	123	124	125	126	127
128	129	131	132	133	134
135	136	137	141	142	143
151					
4	4	4	4	4	4
4	4	4	4	4	4
4	4	4	4	4	4
4	4	4	4	4	4
4					

Les clin

1	2	3	4

Nb de diag poses pour l etude par chaque clin

25	25	25	25

Les diag. diff. poses

1	2	3	4	6	9

Nb de diag. diff. poses

6

Pour chaque patient, on obtient une repartition des cliniciens en groupes de cliniciens en accord entre eux sur un diagnostic, ce qui donne une partition de l'entier 4 :

```
                Diagnostics:----2----
Patient no: 101; Partition:     4
                Diagnostics:----3----
Patient no: 102; Partition:     4
                Diagnostics:----3----
Patient no: 103; Partition:     4
                Diagnostics:----3----6----
Patient no: 104; Partition:     2    2
                Diagnostics:----6----3----
Patient no: 105; Partition:     1    3
                Diagnostics:----3----6----1----
Patient no: 121; Partition:     1    1    2
                Diagnostics:----3----
Patient no: 122; Partition:     4
                Diagnostics:----6----4----
Patient no: 123; Partition:     1    3
                Diagnostics:----2----6----
Patient no: 124; Partition:     2    2
                Diagnostics:----3----
Patient no: 125; Partition:     4
                Diagnostics:----3----
Patient no: 126; Partition:     4
                Diagnostics:----6----2----
Patient no: 127; Partition:     1    3
                Diagnostics:----1----4----
Patient no: 128; Partition:     1    3
                Diagnostics:----4----6----3----
Patient no: 129; Partition:     1    1    2
                Diagnostics:----4----
Patient no: 131; Partition:     4
                Diagnostics:----4----
Patient no: 132; Partition:     4
                Diagnostics:----2----3----4----
Patient no: 133; Partition:     1    1    2
                Diagnostics:----3----
Patient no: 134; Partition:     4
                Diagnostics:----4----6----2----
Patient no: 135; Partition:     1    1    2
                Diagnostics:----2----
Patient no: 136; Partition:     4
                Diagnostics:----6----9----
Patient no: 137; Partition:     2    2
                Diagnostics:----2----
Patient no: 141; Partition:     4
                Diagnostics:----6----2----
Patient no: 142; Partition:     1    3
                Diagnostics:----2----
Patient no: 143; Partition:     4
                Diagnostics:----3----4----
Patient no: 151; Partition:     1    3
```

L'ensemble des partitions possibles est ordonne suivant les valeurs de la variance (non corrigee) de chaque partition (page 143) :

Partition	variance	var corr	concent	concent corr	PAR
1 1 1 1 :	0.000	0.000	0.125	0.000	0.000
2 2 :	0.000	0.000	0.688	0.000	0.333
1 1 2 :	0.222	0.167	0.406	0.305	0.167
1 3 :	1.000	0.889	0.719	0.639	0.500
4 :	2.000	2.000	1.063	1.063	1.000

TABLES:

	LOI BINOMIALE(25,1/4)		LOI BINOMIALE(25,1/2)	
	prob.	prob. cumulees	prob.	prob. cumulees
0	7.5254351E-04	7.5254351E-04	2.9802322E-08	2.9802322E-08
1	6.2711989E-03	7.0237424E-03	7.4505817E-07	7.7486050E-07
2	2.5084795E-02	3.2108538E-02	8.9406967E-06	9.7155571E-06
3	6.4105578E-02	9.6214116E-02	6.8545341E-05	7.8260899E-05
4	0.1175269	0.2137410	3.7699941E-04	4.5526031E-04
5	0.1645377	0.3782787	1.5833975E-03	2.0386579E-03
6	0.1828196	0.5610983	5.2779908E-03	7.3166490E-03
7	0.1654082	0.7265065	1.4325976E-02	2.1642625E-02
8	0.1240562	0.8505626	3.2233447E-02	5.3876072E-02
9	7.8109428E-02	0.9286721	6.0885403E-02	0.1147615
10	4.1658361E-02	0.9703304	9.7416639E-02	0.2121781
11	1.8935615E-02	0.9892660	0.1328409	0.3450190
12	7.3638493E-03	0.9966299	0.1549810	0.5000001
13	2.4546166E-03	0.9990845	0.1549810	0.6549811
14	7.0131914E-04	0.9997858	0.1328409	0.7878220
15	1.7143357E-04	0.9999573	9.7416639E-02	0.8852386
16	3.5715326E-05	0.9999930	6.0885403E-02	0.9461241
17	6.3027042E-06	0.9999993	3.2233447E-02	0.9783576
18	9.3373387E-07	1.000000	1.4325976E-02	0.9926835
19	1.1466908E-07	1.000000	5.2779913E-03	0.9979615
20	1.1466908E-08	1.000000	1.5833974E-03	0.9995449
21	9.1007202E-10	1.000000	3.7699938E-04	0.9999219
22	5.5155880E-11	1.000000	6.8545341E-05	0.9999905
23	2.3980817E-12	1.000000	8.9406967E-06	0.9999994
24	6.6613381E-14	1.000000	7.4505806E-07	1.000000
25	8.8817842E-16	1.000000	2.9802322E-08	1.000000

Aux 25 patients sont attribuees les variances suivantes

2.000000	2.000000	2.000000	0.0000000E+00	1.000000
0.2222222	2.000000	1.000000	0.0000000E+00	2.000000
2.000000	1.000000	1.000000	0.2222222	2.000000
2.000000	0.2222222	2.000000	0.2222222	2.000000
0.0000000E+00	2.000000	1.000000	2.000000	1.000000

Pour chaque valeur possible de la variance, on a la frequence d'observation de cette variance sur les 25 patients:

0.000 0.222 1.000 2.000

3 4 6 12

D'apres les inegalites de Mallows (page 157; 160), on a:

La prob. pour Fmax d exceder la frequence: 3
est comprise entre les deux bornes: 1.000 et 22.595
La prob. pour Fmax d exceder la frequence: 4
est comprise entre les deux bornes: 1.000 et 19.656
La prob. pour Fmax d exceder la frequence: 6
est comprise entre les deux bornes: 1.000 et 10.973
La prob. pour Fmax d exceder la frequence: 12
est comprise entre les deux bornes: 0.081 et 0.084

Conclusion (page 158):
on met dans la classe I la frequence 12 au niveau alpha=.084 ;
on met dans la classe II les autres frequences.

F excede 12 avec probabilite 0.500
I
F excede 18 avec probabilite 0.007
I

Conclusion (page 159):
au niveau alpha=.05 , on postulera le desaccord si les classes I et II sont celles ci-dessus, et on postulera l'accord si l'on rajoute dans la classe I la frequence 6 .

****
****

NOTA BENE

Par la suite on pourra repeter l'analyse que l'on vient d'operer pour la variance corrigee, pour la concentration, pour la concentration corrigee, et pour le PAR.

****
****

Aux 25 patients sont attribuees les variances corrigees suivantes

2.000000	2.000000	2.000000	0.0000000E+00	0.8888889
0.1666667	2.000000	0.8888889	0.0000000E+00	2.000000
2.000000	0.8888889	0.8888889	0.1666667	2.000000
2.000000	0.1666667	2.000000	0.1666667	2.000000
0.0000000E+00	2.000000	0.8888889	2.000000	0.8888889

0.000 0.167 0.889 2.000

3 4 6 12

La prob. pour Fmax d exceder la frequence: 3
est comprise entre les deux bornes: 1.000 et 22.595
La prob. pour Fmax d exceder la frequence: 4
est comprise entre les deux bornes: 1.000 et 19.656
La prob. pour Fmax d exceder la frequence: 6
est comprise entre les deux bornes: 1.000 et 10.973
La prob. pour Fmax d exceder la frequence: 12
est comprise entre les deux bornes: 0.081 et 0.084

F excede 12 avec probabilite 0.500
I
F excede 18 avec probabilite 0.007
I

Aux 25 patients sont attribuees les concentrations suivantes

1.062500	1.062500	1.062500	0.6875000	0.7187500
0.4062500	1.062500	0.7187500	0.6875000	1.062500
1.062500	0.7187500	0.7187500	0.4062500	1.062500
1.062500	0.4062500	1.062500	0.4062500	1.062500
0.6875000	1.062500	0.7187500	1.062500	0.7187500

0.406 0.688 0.719 1.063

4 3 6 12

La prob. pour Fmax d exceder la frequence: 4
est comprise entre les deux bornes: 1.000 et 19.656
La prob. pour Fmax d exceder la frequence: 3
est comprise entre les deux bornes: 1.000 et 22.595
La prob. pour Fmax d exceder la frequence: 6
est comprise entre les deux bornes: 1.000 et 10.973
La prob. pour Fmax d exceder la frequence: 12
est comprise entre les deux bornes: 0.081 et 0.084

F excede 12 avec probabilite 0.500
I
F excede 18 avec probabilite 0.007
I

Aux 25 patients sont attribuees les concentr. corrigees suivantes

1.062500	1.062500	1.062500	0.0000000E+00	0.6388889
0.3046875	1.062500	0.6388889	0.0000000E+00	1.062500
1.062500	0.6388889	0.6388889	0.3046875	1.062500
1.062500	0.3046875	1.062500	0.3046875	1.062500
0.0000000E+00	1.062500	0.6388889	1.062500	0.6388889

0.000 0.305 0.639 1.063

3 4 6 12

La prob. pour Fmax d exceder la frequence: 3
est comprise entre les deux bornes: 1.000 et 22.595
La prob. pour Fmax d exceder la frequence: 4
est comprise entre les deux bornes: 1.000 et 19.656
La prob. pour Fmax d exceder la frequence: 6
est comprise entre les deux bornes: 1.000 et 10.973
La prob. pour Fmax d exceder la frequence: 12
est comprise entre les deux bornes: 0.081 et 0.084

F excede 12 avec probabilite 0.500
I
F excede 18 avec probabilite 0.007
I

Aux 25 patients sont attribuees les PAR suivants

1.000000	1.000000	1.000000	0.3333333	0.5000000
0.1666667	1.000000	0.5000000	0.3333333	1.000000
1.000000	0.5000000	0.5000000	0.1666667	1.000000
1.000000	0.1666667	1.000000	0.1666667	1.000000
0.3333333	1.000000	0.5000000	1.000000	0.5000000

0.167 0.333 0.500 1.000

4 3 6 12

++++++++++++++

Indice d'accord pour la variance: 0.6933334

Indice d'accord pour la variance corr.: 0.6933334

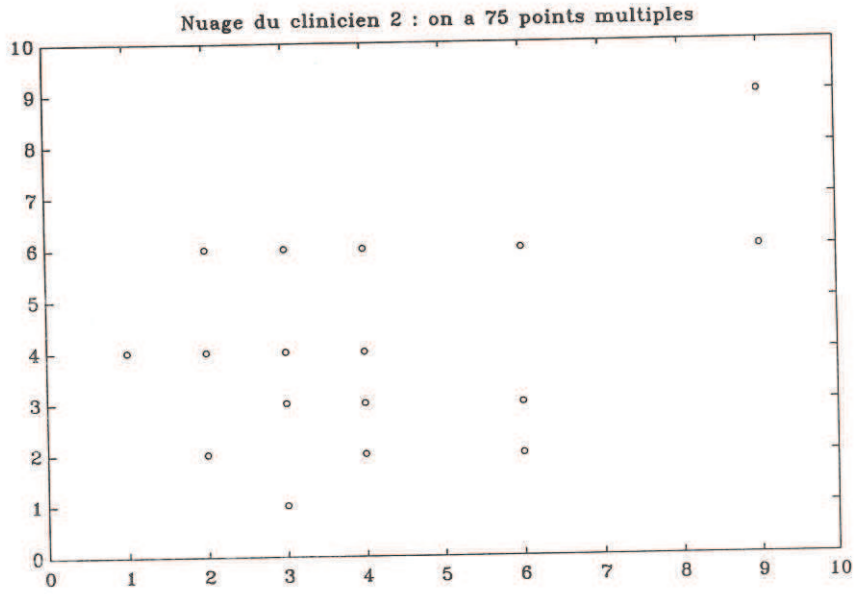
Indice d'accord pour la concentration: 0.6800000

Indice d'accord pour la concentration corr.: 0.6933334

Indice d'accord pour le PAR: 0.6666667

Coefficient de concentration moyen (page 150): 0.7120833

Le nombre de patients pour les calculs : 25



Concentration d'un nuage de points obtenu par simulation

En dimension 2 on considere la matrice de donnees X' :

-8.000000	0.1250000
-7.000000	0.1428571
-6.000000	0.1666667
-5.000000	0.2000000
-4.000000	0.2500000
-3.000000	0.3333333
-2.000000	0.5000000
-1.000000	1.000000
-0.8000000	1.250000
-0.6000000	1.666667
-0.4000000	2.500000
-4.000000	-5.000000
-3.000000	-4.000000
-2.000000	-3.000000
-1.000000	-2.000000
0.0000000E+00	-1.000000
1.000000	0.0000000E+00
2.000000	1.000000
3.000000	2.000000
4.000000	3.000000
5.000000	4.000000
2.000000	0.6931472
3.000000	1.098612
4.000000	1.386294
5.000000	1.609438
6.000000	1.791759
7.000000	1.945910
8.000000	2.079442
9.000000	2.197225
10.00000	2.302585
11.00000	2.397895
12.00000	2.484907
13.00000	2.564949
14.00000	2.639057
15.00000	2.708050

Le nuage contient donc 35 points du plan.

A partir de la, on donne une suite de valeurs de la concentration D en fonction d'une suite de projections definies par une suite de pentes P de droites de projection (on fait 100 projections):

P	D
0.0000000000000000E+00	0.3426005830903790
-0.2030585608340292	0.3419941690962099
-0.2916667498204589	0.3419941690962099
-0.3629515421477292	0.3419475218658892
-0.4259981658247488	0.3420394835485214
-0.4843221723649781	0.3704709704289893
-0.5397428351019730	0.3704723032069971
-0.5933651166166455	0.3990190753852561
-0.6459361016842965	0.3987638483965015
-0.6980043123046792	0.4557201166180758
-0.7499999689559142	0.4844288213244481
-0.8022813466505012	0.5125584339858392
-0.8551619627346185	0.5122785506039150
-0.9089281752317559	0.5117421074552270
-0.9638527856342451	0.5681399416909621
-1.020204095287018	0.5691608496459808
-1.078253083598823	0.6250015826738859
-1.138281130222829	0.5970845481049563
-1.200585839971207	0.5975030403998342
-1.265486774249994	0.6262157434402332
-1.333333250549107	0.6263090379008746
-1.404511005769340	0.6833586005830904
-1.479451074909819	0.6551123698458975
-1.558638832015447	0.6550410662224074
-1.642627247904477	0.6834052478134111
-1.732050807568877	0.6834985422740525
-1.827642588922547	0.6835425239483549
-1.930258625548384	0.7403148688046648
-2.040904042863275	0.7402435651812345
-2.160771572803665	0.7689536026655560
-2.291288010062927	0.7685597667638484
-2.434174999878296	0.7687436901291130
-2.591534057175320	0.7966647230320700
-2.765957404889834	0.7966867138692212
-2.960679894307635	0.7968019991670137
-3.179797199206168	0.7964754685547688
-3.428571808703129	0.7961022907122032
-3.713879439815557	0.7955918367346939
-4.044886862617303	0.7958430653894211
-4.434089720672352	0.7672016659725115
-4.898979789735053	0.7663400249895877
-5.464814276624440	0.7659668471470221
-6.169480301801583	0.7369922532278217
-7.072511476818923	0.7348944606413994
-8.273115430446754	0.7361872553102874
-9.949871974869889	0.7351250312369846
-12.45993841046578	0.7378772178259068
-16.63663895448791	0.7373201166180758
-24.97997857181398	0.7374407330279051
-49.99004669310085	0.7390280716368509
pente verticale	0.7388440000000000
49.99004669310085	0.7102040816326531
24.98001585457065	0.7108791336942940
16.63665554123185	0.7109724281549356
12.45992906729695	0.7109970845481050
9.949871974869889	0.7116488129945856
8.273115430446754	0.7116501457725948
7.072511476818923	0.7118120783007080
6.169482660495137	0.6835438567263640
5.464816146821076	0.6836384839650146
4.898978268891924	0.6836604748021657
4.434089720672352	0.6553002915451895
4.044886862617303	0.6269154518950437
3.713879439815557	0.6269154518950437
3.428571808703129	0.5984839650145773
3.179797893457861	0.5985306122448980
2.960679279925375	0.5986705539358601
2.765956856615527	0.5701457725947522
2.591534057175320	0.5701924198250729
2.434174999878296	0.5701677634319034
2.291288010062927	0.5418295710120783
2.160771945129677	0.5417609329446064
2.040903700016994	0.5418062473969180
1.930258308306063	0.5132361516034985
1.827642588922547	0.5131428571428572
1.732050807568877	0.5131428571428572
1.642627505970596	0.4563005414410662
1.558639074873953	0.4562785506039150
1.479451304321457	0.4562798833819242
1.404511005769340	0.4278017492711370
1.333333250549107	0.4278937109537693
1.265486774249994	0.4278004164931278
1.200585839971207	0.4278017492711370

```
  1.138281312360062       0.3992769679300292
  1.078253259404407       0.3993236151603499
  1.020203924953968       0.3992769679300292
 0.9638527856342451       0.3708441482715535
 0.9089281752317559       0.3707988338192420
 0.8551619627346185       0.3708441482715535
 0.8022815032064436       0.3707988338192420
 0.7500001241763441       0.3422274052478135
 0.6980041574299881       0.3422740524781341
 0.6459361016842965       0.3137492711370262
 0.5933651166166455       0.3137492711370262
 0.5397428351019730       0.3137492711370262
 0.4843221723649781       0.3137959183673484
 0.4259983455085059       0.3137492711370262
 0.3629513444289239       0.3137492711370262
 0.2916665188376579       0.2851311953352770
 0.2030585608340292       0.3422740524781341
0.0000000000000000E+00    0.3426005830903790

 Pour la projection de plus forte concentration on a:
 ####################################################
 P=-2.960679894307635  ;   D=0.7968019991670137

 ----------------------------------------------------
 La droite de concentration minimum D1:
 ----------------------------------------------------
    la pente                      un point
 0.2916665188376579  ;    5.000000       1.609438

 ----------------------------------------------------
 La droite de concentration maximum D2:
 ----------------------------------------------------
    la pente                      un point
 -2.960679894307635  ;    2.000000       0.6931472

 ----------------------------------------------------
 La droite de concentration orthogonale a la droite
 de concentration maximum, soit D3:
 ----------------------------------------------------
    la pente                      un point
 0.3377602563258036       2.000000       0.6931472
```

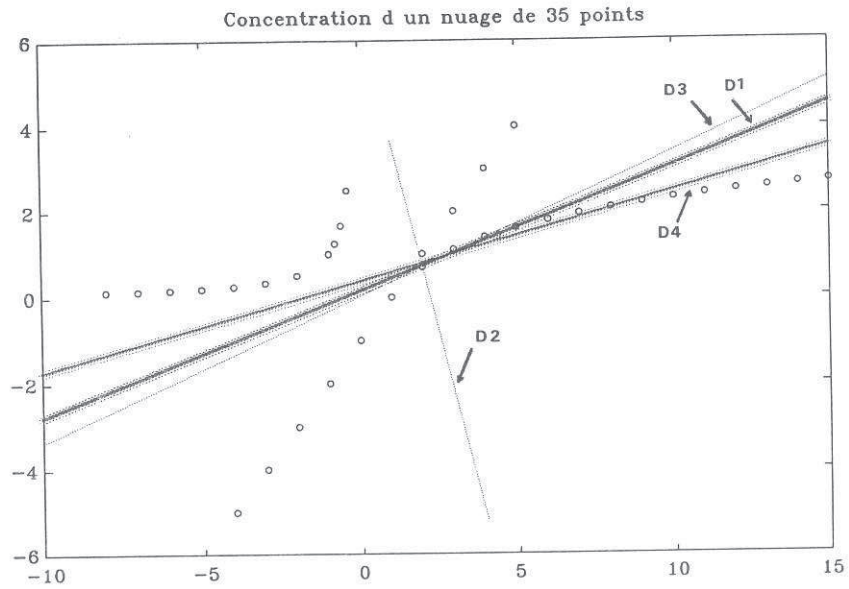
1) On represente le nuage et les trois droites ci-dessus, ainsi que la droite d'inertie minimum D4, comparable a D1, obtenue a l'aide d'une analyse en composantes principales classique (pour D4, la part d'inertie est de 90.2%).

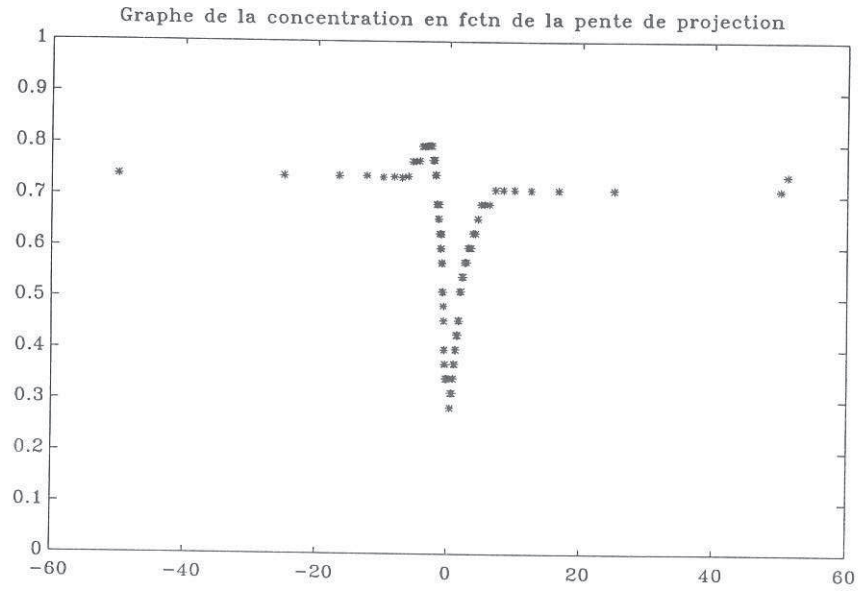
Les calculs ont ete faits en prenant:

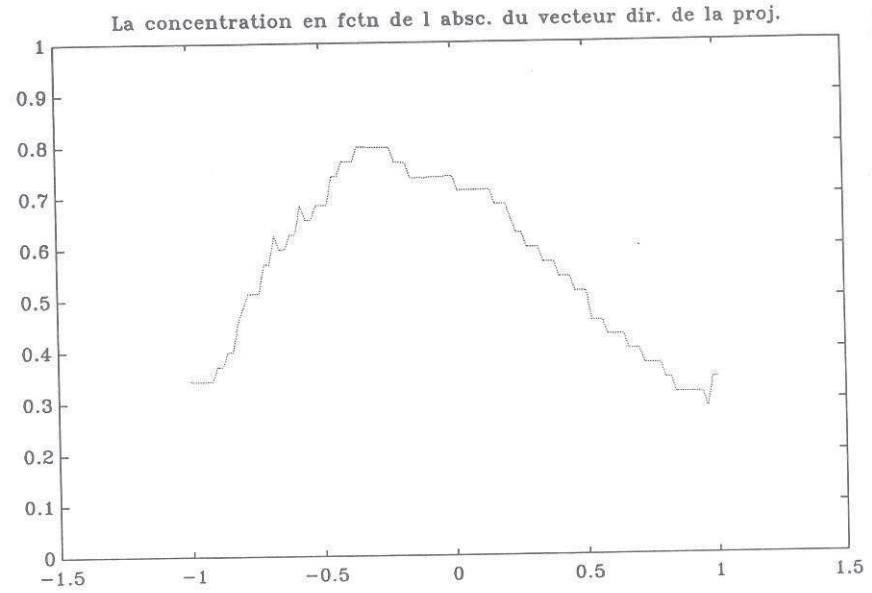
M=3 et N=10

2) Ensuite on donne la concentration en fonction de la pente de projection (D en fonction de P) ; on a 100 points.

3) Enfin on donne la concentration en fonction de l'abscisse du vecteur directeur norme de la droite de projection. Cette abscisse varie entre -1 et 1 par pas de .02. Le graphe donne bien l'idee des minima et maxima locaux de D.




Graphe de la concentration en fctn de la pente de projection
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
−60
−40
−20
0
20
40
60


La concentration en fctn de l absc. du vecteur dir. de la proj.
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
−1.5
−1
−0.5
0
0.5
1
1.5

B I B L I O G R A P H I E

Andrews, G.E. (1976), "The Theory of Partitions", Addison-Wesley Publishing Company.

Bartko, J.J. (1966), "The Intraclass Correlation Coefficient as a Measure of Reliability", *Psychol. Reports*, 19, pp. 3-11.

Beckett, J., Schucany, W.R. (1975) "ANACONDA: Analysis of Concordance of g Groups of Judges", Proceedings of Soc. Stat. Section of the Am. Stat. Ass., pp. 311-313.

Caillez, F., Pages, J.-P. (1976) "Introduction à l'Analyse des Données", SMAC, Paris.

Cartwright, D.S. (1956) "A Rapid Non-Parametric Estimate of Multi-Judge Reliability", *Psychometrika*, 21, pp. 17-29.

Chandon, J.-L., Pinson, S. (1981) "Analyse Typologique, Théories et Applications", Masson, Paris.

Cohen, J. (1960) "A Coefficient of Agreement for Nominal Scales", *Educ. and Psychol. Measurement*, 20, pp. 37-46.

Critchley, F. (1988) "L'Analyse des Dissimilarités: Quelques Progrès Récents, Problèmes Courants et Perspectives pour l'Avenir", Communication aux XXèmes Journées de statist., Grenoble.

De La Vallée Poussin, C. (1950) "Intégrales de Lebesgue, Fonctions d'Ensemble , Classes de Baire", Gauthier-Villars, Paris.

Diday, E., Lemaire, J., Pouget, J., Testu, F. (1982) "Ele ments d'Analyse de Données", Dunod, Paris.

Ellison, W.J., Mendés-France, M. (1975) "Les Nombres Premiers", Hermann, Paris.

Fleiss, J.L. (1971) "Measuring Nominal Scale Agreement among Many Raters", *Psychol. Bulletin*, 76, 5, pp. 378-382.

Friedman, J.H. (1987) "Exploratory Projection Pursuit", *J. Am. Stat. Ass.*, 82, pp. 249-266.

Gordon, A.D. (1979) "A Measure of Agreement Between Rankings", *Biometrika*, 66, pp. 7-15.

Guiard, V. (1984) "Systems of One-Dimensional Continous Distributions and their Application in Simulation Studies", D. Reidel Publishing Company.

Hermann, P., Hovaguimian, T. (1983) "Une Méthode d'Evaluation des Interventions du Thérapeute dans le Cadre d'Entretiens à Visée Psychotérapique", *An. Med. Psychol.*, 141, 6, pp. 601-605.

Hollander, M., Sethuraman, J. (1978) "Testing for Agreement between two Groups of Judges", *Biometrika*, 62, 2, pp. 403-411.

Huber, P.J. (1985) "Projection Pursuit", *The Annals of Statistics*, 13, pp. 435-475.

Huber, P.J. (1987) "Projection Pursuit", Communication aux XIXèmes Journées de statist., Lausanne.

Johnson, N.L., Kotz, S. (1969) "Discrete Distributions", Houghton Mifflin Company, Boston.

Jones, M.C., Sibson, R. (1987) "What is Projection Pursuit", *J. R. Statist. Soc. A*, 150, Part I, pp. 1-36.

Kendall, M.G. (1970) "Rank Correlation Methods", 4th ed., Griffin, London.

Kendall, M.G., Stuart, A. (1976) "The Advanced Theory of Statistics", 4th ed., Griffin, London.

Landis, J.R., Koch, G.G. (1975) "A Review of Statistical Methods in the Analysis of Data arising from Observer Reliability Studies (Part I)", *Stat. Neerlandica*, 29, 3, pp. 101-124.

Launer, R.L., Siegel, A.F. (1982) "Modern Data Analysis", Academic Press Inc., London.

Lecoutre, J.P., Tassi, P. (1987) "Statistique non Paramétrique et Robustesse", Economica, Paris.

Mallows, C.L. (1968) "An Inequality Involving Multinomial Probabilities", *Biometrika*, 55, pp. 422-424.

Marshall, A., Olkin, I. (1979) "Inequalities: Theory of Majorization and its Applications", Academic Press.

Raffestin, C., Tricot, C. (1974) "Réflexions sur les Formes", Actes du 3e Colloque sur l'Analyse des Données en Géographie, Besançon.

Renyi, A. (1966) "Calcul des Probabilités", Dunod, Paris.

Robinson, W.S. (1957) "The Statistical Measurement of Agreement", *Am. Soc. Rev.*, 22, pp. 17-25.

Schouten, H.J.A. (1980) "Measuring Pairwise Agreement among many Observers", *Biometrical Journal*, 22, 6, pp. 497-504.

Schouten, H.J.A. (1982) "Measuring Pairwise Agreement among many Observers. II. Some Improvements and Additions", *Biometrical Journal*, 24, 5, pp. 431-435.

Sibson, R. (1972) "Order Invariant Methods for Data Analysis", *J. R. Statist. Soc. B*, 34, pp. 311-338.

Tricot, C. (1971) "Concentration d'un Ensemble de Points", Cahiers du Département d'Econométrie de l'Université de Genève.

Tricot, C. (1988) "Des Réseaux", Rendiconti del Seminario Matematico di Brescia, 10, Università Cattolica, Milano.

Verducci, J.S., Mack, M.E., DeGroot, M.H. (1988) "Estimating Multiple Rater Agreement for a Rare Diagnosis", *J. of Multivariate Analysis*, 27, pp. 512-535.