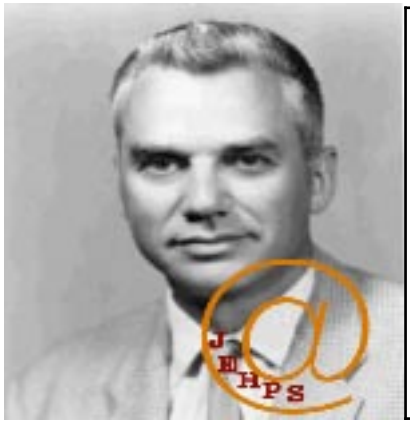

Journ@l Electronique d'Histoire des Probabilités et de la Statistique

Electronic Journ@l for History of Probability and Statistics

Vol 4, n°2; Décembre/December 2008

www.jehps.net

L'analyse des données des origines à 1980 : quelques éléments.

Ludovic LEBART[1]

Ce numéro est dédié à la mémoire de Henry Rouanet, décédé pendant la préparation de ce recueil de contributions auquel il était invité à participer. Ingénieur, mathématicien et psychologue, Directeur de Recherches au CNRS, Henry Rouanet fut un pionnier de la statistique multidimensionnelle et de l' « analyse géométrique des données ». On trouvera des textes historiques d'Henry Rouanet relatifs au thème de ce numéro spécial dans la section « Documents » du présent recueil.

Ce dossiers thématique du journal est consacré aux origines et au développement des méthodes exploratoires statistiques multidimensionnelles durant le vingtième siècle, sans aller au delà des années 1980. Notons d'emblée que l'équivalent anglais de « analyse des données » serait à peu près *exploratory multivariate data analysis*. . L'expression anglaise *data analysis* a en effet un sens plus général de statistique appliquée (avec une connotation d'approche pragmatique et informatisée).

Après 1980 eurent lieu diverses explosions et apparurent de nouveaux paradigmes qui ne sont pas encore stabilisés, et pour lesquels les contributions présentées ici constituent des matériaux qui pourront être retravaillés ou refondus par la suite. Ces années ont en effet vu l'apparition des méthodes neuronales (*neural networks*), des cartes auto-organisées (*self organising maps*), du *data mining* (fouilles de données), de la théorie de l'apprentissage (*learning theory*), de l'analyse en composantes indépendantes, des méthodes de rééchantillonnage, autant de méthodes, d'écoles ou de courants qui ont vocation à rejoindre le thème qui nous intéresse, mais donnant lieu encore à des débats, voire des controverses, et à une dispersion terminologique importante. Plusieurs auteurs de ce numéro thématique ont d'ailleurs été ou sont encore des acteurs importants de l'analyse exploratoire

[1] TELECOM-ParisTech. 46 rue Barrault, 75013, Paris, France. ludovic@lebart.org

multidimensionnelle, ce qui donne de toutes façons à ce recueil de témoignages un intérêt documentaire certain.

Plusieurs pays ou Ecoles sont représentés par des articles ou des documents mais des lacunes importantes subsistent. Malgré l'internationalisation et la globalisation de l'activité scientifique, les lieux de production, comme les lieux d'observation restent des facteurs importants d'explication et d'interprétation de cette activité. Nous verrons que les barrières linguistiques, et même politiques, ont eu des répercussions sur la communication scientifique.

Ce numéro du Journal comporte neuf contributions originales et trois séries de documents d'archives. Tous les auteurs sont statisticiens. Certains d'entre eux ont cependant choisi d'intervenir en tant que sociologues ou historiens.

1 Description, exploration, confirmation

La statistique descriptive classique permet de représenter sous forme de graphiques des informations statistiques en les simplifiant et les schématisant. La *statistique descriptive multidimensionnelle* en est la généralisation naturelle lorsque ces informations concernent plusieurs variables ou dimensions.

Mais le multidimensionnel induit un changement qualitatif important. Pour reprendre une analogie répandue, les microscopes ou les appareils radiographiques ne sont pas seulement des instruments de description. Ce sont aussi des instruments d'observation ou d'exploration, et des outils de recherche. De la même façon, la réalité multidimensionnelle n'est pas seulement simplifiée parce que complexe, mais aussi explorée parce que cachée. Il faut préparer et coder les données, utiliser des règles d'interprétation strictes et valider les représentations fournies par les techniques utilisées dans le cas multidimensionnel. Ces opérations n'ont pas la simplicité de la statistique descriptive élémentaire. Il ne s'agit pas seulement de présenter mais d'analyser, de découvrir, parfois de vérifier et prouver.

Les nouveaux outils de calcul

Née avec le vingtième siècle à la suite des travaux de précurseurs comme l'astronome Quetelet et les démographes, biométriciens et statisticiens Galton, Pearson, puis Fisher, la science statistique a manipulé des chiffres pendant un demi-siècle sans disposer de véritables outils de calcul. Les appareils que l'on trouve maintenant dans la poche des écoliers et dans la plupart des foyers auraient comblé les aspirations les plus insensées des statisticiens jusqu'en 1960.

Face à ces nouvelles possibilités, John W. Tukey, le fondateur du courant désigné par *Exploratory Data Analysis* (*EDA*), a une attitude assez novatrice (cf. Tukey, 1977, et les publications antérieures de cet auteur cités par les différents auteurs de ce numéro thématique). Dans une optique plus spécifiquement multidimensionnelle, (cf. J.-P. Benzécri, 1973 et les autres publications de cet auteur également citées) affirme que l'ordinateur impose de repenser entièrement la statistique.

Ces deux pionniers n'ont pas eu l'influence immédiate que l'on aurait pu attendre en dehors de leurs aires d'influence directes. A défaut d'être repensée, la statistique s'est cependant progressivement enrichie. La période récente a connu des changements tout à fait notables du fait de la diffusion des moyens de calcul : les outils existants ont été améliorés, de nouveaux outils sont apparus, de nouveaux domaines d'application ont été explorés.

Il est possible maintenant de traiter des tableaux correspondant à des dizaines de milliers d'observations et des centaines, voire des milliers de variables. Le changement d'échelle des données a rapidement conduit à modifier les outils eux-mêmes et à imaginer de nouveaux outils et de nouvelles approches. Mais les statisticiens savent qu'il est parfois vain de vouloir traiter des millions d'observations lorsqu'il y a des possibilités d'échantillonnage.

La levée de l'obstacle du calcul a eu pour effet de diffuser l'emploi des techniques de type algorithmique, au premier rang desquelles se trouvent les techniques de classification automatique et les méthodes impliquant des algorithmes coûteux. D'autres techniques, comme les techniques de sélection pas-à-pas, les techniques d'estimation par la méthode du maximum de vraisemblance, de programmation dynamique, de recherches automatiques de règles dans des bases de données connaissent des utilisations à grande échelle.

L'étude statistique des variables qualitatives est par nature plus complexe que celle des variables numériques continues, qui s'appuie généralement sur la loi normale et sur les formalismes simples qui en dérivent (maximum de vraisemblance, moindres carrés, par exemple). Il n'est donc pas étonnant que les possibilités de calcul aient permis de fortes avancées dans ce domaine : analyse des correspondances simples et multiples dans le cas descriptif, modèles log-linéaires, discrimination et modèles logistiques dans le cas inférentiel.

Une des innovations de la statistique après 1960 aura été la matérialisation des techniques sous forme de "produits", les logiciels, développés avec des contraintes économiques et commerciales de conception, de production, de distribution. Comme tout produit fini, le logiciel a l'avantage de diffuser et l'inconvénient de figer. Comme tout produit à l'usage de spécialistes, il induit de nouvelles divisions du travail, parfois peu souhaitables dans un processus de connaissance. Les logiciels accessibles et faciles à utiliser permettront une large diffusion des méthodes, mais donneront lieu à des utilisations inconsidérées dans des domaines où une réflexion minutieuse et une grande prudence seraient de mise.

On distingue deux types d'approches en statistique multidimensionnelle : les *approches descriptives et exploratoires* (qui sont souvent les approches non-supervisées de la théorie de l'apprentissage) et les *approches inférentielles et confirmatoires* (dont font partie les approches supervisées) qui constituent le volet le plus ample et le plus classique de la science statistique.

Rappelons brièvement les caractéristiques de ces deux familles de méthodes, qui correspondent à des approches complémentaires.

- *La statistique descriptive et exploratoire* permet, par des résumés et des graphiques plus ou moins élaborés, de décrire des ensembles de données statistiques, d'établir des relations entre les variables sans faire jouer de rôle privilégié à une variable particulière. C'est à cette famille de méthode que ce numéro est consacré. Classiquement, les conclusions ne portent dans cette phase de travail que sur les données étudiées, sans être inférées à une population plus large. L'analyse exploratoire s'appuie essentiellement sur des représentations graphiques et sur les techniques descriptives multidimensionnelles (analyse en composantes principales, analyse des correspondances, classification). Les méthodes de ré-échantillonnage actuelles permettent de valider des structures et donc d'articuler, avec prudence, exploration et inférence.

- *La statistique inférentielle et confirmatoire* permet de valider ou d'infirmer, à partir de tests statistiques ou de modèles probabilistes, des hypothèses formulées *a priori* (ou après une phase exploratoire séparée), et d'extrapoler, c'est-à-dire d'étendre certaines propriétés d'un échantillon à une population plus large. Les conclusions obtenues à partir des données vont au delà de ces données. La statistique confirmatoire fait surtout appel aux méthodes dites

explicatives et prévisionnelles destinées, comme leurs noms l'indiquent, à expliquer puis à prévoir, suivant des règles de décision, une variable privilégiée à l'aide d'une ou de plusieurs variables explicatives (régressions multiples et logistiques, analyse de la variance, analyse discriminante, segmentation, etc.).

Les démarches sont complémentaires, l'exploration et la description devant en général précéder les phases explicatives et prédictives. En effet, une exploration préliminaire est souvent utile pour avoir une première idée de la nature des liaisons entre variables, et pour traiter avec prudence les variables corrélées et donc redondantes qui risquent de charger inutilement les modèles.

Cependant, les démarches elles-mêmes ne sont pas toujours faciles à discerner, à identifier. L'exploration pure est rare, car il existe toujours des informations et des connaissances *a priori* sur le tableau de données *(meta-data)*, et donc des hypothèses générales, ou même des attentes non explicitement formulées de la part de l'utilisateur. Cox (1977) dans un remarquable article de synthèse sur les tests de signification, met en évidence les problèmes que pose l'articulation *exploratoire – confirmatoire*. On ne peut évidemment pas tester statistiquement sur des données de nouvelles hypothèses suscitées par ces mêmes données. Mais il serait absurde et réducteur de nier que des recueils de données peuvent aussi suggérer des hypothèses (surtout les vastes recueils de données multidimensionnelles). Il reste alors à respecter une discipline de travail (fragmentation, réplication, reproduction et renouvellement du recueil) contraignante mais productive.

Quatre sections vont structurer la suite de ce chapitre introductif : Les sections 2 et 3 dévolues respectivement aux analyses en axes principaux et aux méthodes de classification correspondent en fait à la quasi totalité des neuf contributions qui constituent ce numéro. Les sections 4 et 5 dévolues aux méthodes de validation et aux thèmes connexes évoquent des domaines de recherche non abordés ou simplement évoqués par les contributeurs à ce numéro spécial.

2 Méthodes factorielles (*principal axes methods*)

- Les *méthodes factorielles*, ou encore *analyses en axes principaux*, opèrent une réduction de certaines représentations "multidimensionnelles" et produisent essentiellement des *visualisations graphiques* planes ou parfois tridimensionnelles des éléments à décrire. Les techniques d'analyse factorielle comprennent dans la littérature statistique française toutes les techniques de représentation utilisant des "axes principaux": analyse en composantes principales, des correspondances simples et multiples, analyse factorielle dite classique ou des psychologues — alors que l'expression correspondante en anglais (*factor analysis*) ne désigne de façon assez stricte que cette dernière technique : analyse en facteurs communs et spécifiques de Spearman, Thurstone, utilisée principalement par les psychologues et les psychométriciens.

A la base de ces analyses en axes principaux se trouve un théorème, la *décomposition aux valeurs singulières (Singular Values Decomposition)*, présenté pour la première fois par Eckart et Young (1936) pour les tableaux rectangulaires, généralisant les travaux de Sylvester (1889) relatifs aux matrices carrées. Gifi (1981/1990) mentionne également les travaux antérieurs et indépendants de Beltrami (1873) et Jordan (1874). Le problème que l'on se propose de résoudre est alors un problème de réduction purement numérique, autrement dit,

un problème de compression de données : l'ajustement au sens des moindres carrés d'une matrice par une matrice de rang inférieur. On notera également, parmi les premiers articles publiés sur les bases algébriques et géométriques des méthodes en axes principaux : Gower (1966), Gabriel (1971).

La contribution de **John C. Gower** à ce numéro spécial du JEHPS **[The biological stimulus to multidimensional data analysis]** couvre aussi bien l'histoire des méthodes en axes principaux que celle des méthode de classification qui sont évoquées en section 3. Pour ces deux familles de méthodes, J. Gower a été un témoin privilégié et un créateur. Sa contribution est présentée à propos des méthodes factorielles car il a semblé, a priori, que la majorité de ses travaux se situaient dans ce cadre. Pourtant, le récit de la genèse des méthodes de classification, de leurs premiers balbutiements, des incompréhensions sur leur portée, de leur accueil par différentes communautés sera probablement considéré comme un des points fort de ce recueil. Comme le titre l'indique, l'emphase est mise sur le rôle de la biométrie, de l'anthropométrie, de l'agronomie et des sciences naturelles dans le développement des analyses multidimensionnelles, après les travaux fondateurs de K. Pearson et R. Fisher.

2.1 L'analyse en composantes principales

L'analyse en composantes principales est la technique de visualisation en axes principaux la plus ancienne et la plus répandue. Conçue pour la première fois dans un cadre limité par Karl Pearson en 1901, intégrée à la statistique mathématique par Harold Hotelling en 1933, l'analyse en composantes principales n'est vraiment utilisée que depuis l'avènement et la diffusion des moyens de calculs. Pour le statisticien classique, il s'agit de la recherche des axes principaux de l'ellipsoïde indicateur d'une distribution normale multidimensionnelle, ces axes étant estimés à partir d'un échantillon. C'est la présentation initiale de Hotelling (l933), puis celle des manuels classiques d'analyse multivariée, comme l'ouvrage fondamental d'Anderson (1958). Pour les factorialistes classiques, il s'agit d'un cas particulier de la méthode d'analyse factorielle des psychométriciens (cas de variances spécifiques nulles ou égales ; cf. Horst, 1965; Harman, 1967). Enfin, du point de vue plus récent des analystes de données, il s'agit d'une technique de représentation des données, ayant un caractère optimal selon certains critères algébriques et géométriques et que l'on utilise en général sans référence à des hypothèses de nature statistique ni à un modèle particulier. Ce point de vue, fort répandu actuellement est peut-être le plus ancien. C'est celui qui avait été adopté par Pearson (1901). On trouvera une présentation plus proche des préoccupations actuelles dans l'article de synthèse de Rao (1964).

2.2 L'analyse factorielle classique *(factor analysis)*

A l'origine des principes de la méthode principalement par les psychologues et psychométriciens se trouvent Spearman (1904) (analyse monofactorielle), puis Garnett (1919) et Thurstone (1947) (analyse multifactorielle).Bien qu'il s'agisse d'un modèle statistique particulier, et non d'une méthode exploratoire, ses liens sont profonds avec l'analyse en composantes principales. Les développements auxquels il donne lieu sont complexes et diversifiés. On pourra consulter sur ce point les ouvrages de Harman (1967), Mulaik (1972). Mentionnons pour conclure ce bref aperçu les travaux historiques d'Anderson et Rubin (1956) et de Lawley et Maxwell (1963) qui ont placé l'analyse factorielle en facteurs communs et spécifiques dans un cadre inférentiel classique.

2.3 L'analyse des correspondances

L'analyse des correspondances constitue l'autre technique factorielle fondamentale. La plupart des autres techniques dérivent de ces deux techniques de base pour s'adapter à des domaines d'application spécifiques. L'une des plus utilisées est l'analyse des correspondances multiples applicable aux grands fichiers de variables nominales.
L'analyse des correspondances, présentée sous ce nom pour la première fois et développée par Escofier-Cordier (1965) et Benzécri (1969), a un certain nombre de précurseurs, parmi lesquels il faut citer Guttman (1941), Hayashi (1956). Ces deux auteurs ont proposés, indépendamment, la technique comme étant destinée à analyser des données.
La contribution de **Fionn Murtagh**, dans le présent numéro du JEHPS**, [Origins of Modern Data Analysis Linked to the Beginnings and Early Development of Computer Science and Information Engineering]** analyse de façon plus détaillée la spécificité de la démarche de Chikio Hayashi, fondateur par ailleurs de concept de « Data Science ». Le document annexe numéro 13 contient une biographie du Professeur Hayashi. Nous reviendrons sur la contribution de F. Murtagh plus bas. L'analyse des correspondances pouvant être présentée selon divers points de vue, il est difficile d'en faire un historique précis. Les principes théoriques remontent probablement aux travaux de Fisher (1940) sur les tables de contingences, dans un cadre de statistique inférentielle classique. Depuis les travaux de Benzécri (1973) et de Escofier-Cordier (1965), on utilise surtout les propriétés algébriques et géométriques de l'outil descriptif que constitue l'analyse. Les ancêtres les plus lointains de l'analyse des correspondances seraient, de façons tout à fait indépendantes, Richardson et Kuder (1933) et Hirschfeld (1935). Richardson et Kuder visaient une meilleure sélection des vendeurs pour la société *Procter and Gamble*, et découvraient empiriquement la méthode de « reciprocal averaging » alors que Hirschfeld découvrait une importante propriété de statistique mathématique (voir sur ce point la contribution de J. Gower). Jan de Leeuw (1983) montre que Karl Pearson était sur le point de découvrir l'analyse des correspondances en 1906, mais il lui manquait (comme à tous ses contemporains) une connaissance des propriétés spectrales des matrices. Cette variété de contextes est caractéristique de l'analyse des correspondances, méthode aussi utile en pratique que stimulante du point de vue théorique. Cf. l'ouvrage de Escofier (2003) qui reprend tous les travaux de Brigitte Escofier-Cordier. Cf. également les références historiques de Hill (1974), puis l'ouvrage historique de Benzécri (1982).

L'article de **Michel Armatte** dans le présent numéro **[Histoire et Préhistoire de l'Analyse des données par J.P. Benzécri : un cas de généalogie rétrospective]** analyse du point de vue d'un historien des sciences et avec le recul d'une trentaine d'année le travail réalisé par J.P. Benzécri après la parution de l'article précité de M.O. Hill (Correspondence Analysis : a neglected multivariate method) en 1974. Ce travail sous forme de recueil de documents polycopiés fut ensuite publié dans les *Cahiers de l'Analyse des données* en 1976/77, puis sous forme d'ouvrage chez Dunod (Paris) en 1982. C'est peut-être le titre de l'article de Hill (« *a neglected method* ») qui était irritant pour une communauté qui voyait au contraire depuis plusieurs années une utilisation frénétique et souvent excessive de l'analyse des correspondances dans les domaines les plus divers (« *an overused method* » aurait été mieux reçu). Cet article d'historien est aussi le témoignage vivant d'un ancien étudiant de Benzécri, ce qui rend sa lecture aussi agréable qu'instructive. La réflexion en forme de boutade de Michel Armatte « L'histoire des sciences est une discipline bien trop importante pour la laisser aux scientifiques…que l'on veut étudier! » est de toute façon bien à sa place dans ce numéro

du JEHPS auquel participent essentiellement des scientifique profondément et souvent passionnément impliqués dans leur discipline.

La contribution d'**Alain Desrosières [Analyse des données et sciences humaines : comment cartographier le monde social?]** montre précisément l'impact important des méthodes d'analyse des données sur les méthodes des sciences sociales dans les années 70 en France, impact qui confine à l'engouement, attesté, par exemple, par la publications de plans factoriels dans des magazines à grands tirages. De façon assez étonnante, la méthodologie eût elle-même une certaine coloration politique. Même le temple de la statistique officielle (l'INSEE, dont Alain Desrosières fait partie) fut un temps perturbé par la fébrilité des sociologues et des économistes qui découvraient enfin un outil à la mesure de la complexité de leur objet. L'influence s'est encore exercée au cours des années récentes et continue à l'être dans le contexte de l'"analyse géométrique des données" avec les travaux de Henry Rouanet et Brigitte Le Roux, souvent auprès de chercheurs disciples du sociologue Pierre Bourdieu.

La contribution de Fionn Murtagh déjà citée évoque aussi le contexte informatique particulier de l'apparition et du développement de l'analyse des données, mais surtout contient une analyse très fine des obstacles linguistiques et culturels qui freinèrent la compréhension et la diffusion des méthodes. L'absence de traduction des travaux du laboratoire de Benzécri pendant les années 70 a conduit à des incompréhensions avec le monde anglo-saxon, qui contrastent avec l'adhésion de pays proches linguistiquement comme l'Italie (voir la contribution de A. Rizzi dans ce recueil). Le malentendu a été prolongé plus tard lors de la traduction en Anglais d'une version simplifiée du « Traité d'Analyse des Données » de Benzécri à l'usage des praticiens, proche du manuel « Pratique de l'analyse des Données » de Bastin et al. (1980), sous le titre « Correspondence Analysis Handbook » (Benzécri, 1992). Ce dernier ouvrage a été salué par erreur comme étant la traduction du traité de 1973. « This is a translation of Benzécri 'bible' on correspondence analysis previously available only in French » écrit David Hand (1994) dans le « Journal of Classification ». Bien qu'il ne s'agisse pas d'un ouvrage destiné aux statisticiens, D. Hand remarque cependant, à l'actif du livre: « With its many examples of correspondence analysis being applied (in different ways) the book provides an excellent illustration of how sensitive and sophisticated use of a single technique can shed light on data on many different ways. It serves to support the position that a thorough grasp of a few techniques is better than a weak grasp of many.».

2.4 L'analyse des correspondances multiples

L'analyse des correspondances multiples est une simple extension du domaine d'application de l'analyse des correspondances appliquée non plus à une table de contingence, mais à un *tableau disjonctif complet*. Les propriétés d'un tel tableau sont intéressantes, les procédures de calculs et les règles d'interprétation des représentations obtenues sont simples et spécifiques. On peut faire remonter les principes de cette méthode à Guttman (1941), mais aussi à Burt (1950) ou à Hayashi (1956). D'autres types d'extension de l'analyse des correspondances à partir de l'analyse canonique généralisée s'appuient notamment sur les travaux de Carroll (1968), Horst (1961) et Kettenring (1971).

L'analyse des correspondances multiples a été développée également sur le nom d'*Homogeneity Analysis* par l'équipe de J. de Leeuw depuis 1973 (cf. Gifi, 1981/1990) et sous le nom de *Dual Scaling* par Nishisato (1980). Une application de l'analyse des

correspondances à un tableau disjonctif complet se trouve dans Nakache (1973). Un exposé synthétique de ces diverses approches a été réalisée par Tenenhaus et Young (1985).

La contribution de **Willem Heiser** à ce numéro du JEHPS **[Psychometric Roots of Multidimensional Data Analysis in the Netherlands: From Gerard Heymans to John van de Geer]** contient des informations précieuses sur tous les développements des méthodes en axes principaux. Il y a un consensus parmi tous les auteurs de ce numéro spécial sur l'importance fondatrice de la psychologie (et plus précisément de la psychologie différentielle, comme le soulignent à la fois Heiser et Desrosières). W. Heiser décrit comment les deux psychologues Gerard Heymans, puis John van de Geer ont méthodiquement manipulé et enregistré des données multidimensionnelles et comment la "Dutch School of psychology" (reconnue notamment par Spearman et ses contemporains) a finalement favorisé la naissance d'une "Dutch School of data analysis", basée principalement à Leiden et à Groningen. Les lecteurs pourront savoir aussi qui est Gifi (le Bourbaki statistique néerlandais), un des auteurs les plus cité de ce numéro.

La contribution d'**Antoine de Falguerolles [L'analyse des données ; before and around]** donne un autre point de vue, celui du statisticien universitaire (de plus : cultivé et curieux) sur d'une part la situation de l'analyse des données en France pendant la période 1965-1985, qui, on l'a vu, fut caractérisée par une fébrilité exceptionnelle, d'autre part sur plusieurs situations historiques parfois très anciennes (1588 pour la « Felissima Armada »…) ayant conduit à l'examen (par encore l'analyse !) de données multidimensionnelles. Une brève section évoque aussi l'introduction d'une modélisation probabiliste dans certaines démarches d'analyse des données. Cette contribution illustre aussi le regain d'intérêt (lié peut-être à la disponibilité d'outils multimedia de communications) pour les méthodes graphiques et leur histoire.

La contribution de **Alfredo Rizzi [Italian Contributions to Data Analysis]** est présentée dans le cadre des méthodes factorielles car elle consacre une part plus importante à ce thème, mais elle concerne aussi les méthodes de classifications pour lesquelles les chercheurs italiens ont été très actifs. A. Rizzi rappelle tout d'abord que le premier congrès de l'Institut International de Statistique s'est tenu à Rome en 1887 et évoque les travaux de pionniers de Corrado Gini en 1912 (il s'agissait déjà d'une approche descriptive élaborée). Les auteurs qu'ils cite ensuite sont tous connus pour leurs publications en Anglais, mais les textes italiens cités sont soit originaux, soit parus à une date antérieure, ce qui présente un intérêt historique certain.

2.5 Tableaux multiples

Il n'existe pas d'analogue du théorème d'Eckart et Young dans le cas des tableaux tridimensionnels. Ce que l'on peut exprimer dans les termes suivants : il existe une décomposition hiérarchique unique d'un élément du produit tensoriel de deux espaces euclidiens en une somme de produits tensoriels de vecteurs appartenant à chacun des deux espaces. Mais une telle décomposition n'est pas unique dans le cas de d'un élément du produit tensoriel de plus de deux espaces euclidiens (cf. Benzécri, 1973; Tome 2B, n°6). Il ne peut donc exister dans ce cas de démarche exploratoire aussi bien assise que dans le cas des tableaux à double entrée.

Evoquons quelques travaux de référence sur le thème des tableaux à plusieurs dimensions On trouvera une synthèse et une classification des principales démarches dans l'ouvrage de

Kroonenberg (1983). Les premiers travaux sur ce thème sont ceux de Tucker (1964, 1966) puis ceux de Harshman (1970), tous les deux dans le cadre de l'analyse factorielle classique.

L'analyse canonique généralisée a été présentée dans Horst (1961), où elle figure au troisième rang parmi quatre généralisations possibles de l'analyse canonique. Elle a été reprise ou développée par Carroll (1968) et Kettenring (1971). Citons également les travaux de Pagès et al.(1976) fondés sur les opérateurs définis par Robert et Escoufier (1976).

Les méthodes d'analyse procrustéennes tentent de répondre à une préoccupation fréquente en statistique multidimensionnelle : *n* individus ou observations sont décrits d'une part par *p* variables (colonnes de X), d'autre part par *q* autres variables (colonnes de Z). Comment comparer les deux nuages d'individus, les deux systèmes de distances entre individus ? C'est Tucker (1958) qui proposa à l'origine une telle méthode pour comparer deux batteries de tests passés sur les mêmes individus. La technique a ensuite été étudiée par Cliff (1966), Schönemann (1968), Schönemann et Carroll (1970), puis généralisée par Gower (1975).

3 La classification *(Clustering)*

Les *méthodes de classification non supervisées* (appelées souvent *méthodes de classification* dans le monde francophone) produisent des groupements en classes d'objets pour les méthodes de partitionnement, ou en familles de classes hiérarchisées pour les méthodes de classification hiérarchiques. Les éléments à décrire sont groupés de la manière la moins arbitraire possible à partir de leurs vecteurs de description.

La classification est une branche de l'analyse des données qui a donné lieu à des publications nombreuses et diversifiées. Elle s'est beaucoup développée, ces dernières années, pour répondre au besoin d'extraire de façon automatique l'information cachée ou d'identifier des groupes ou classes à partir d'importantes masses d'information de plus en plus spécifiques. L'ouvrage de base, historique, est probablement celui de Sokal et Sneath (1963) suivant l'article séminal de Sneath (1959). Les premiers manuels publiés furent ceux de Lerman (1970), Jardine and Sibson (1971), Anderberg (1973), Benzécri (1973), Bock (1974), Hartigan (1975). Une des premières synthèses historiques sur le sujet est celle de Cormack (1971). Une synthèse de travaux plus récents en classification hiérarchique a été faite par Gordon (1987).

3.1 Centres mobiles et k-means

La contribution de **Hans Hermann Bock** dans le présent numéro **[Origins and extensions of the k-means algorithm in cluster analysis]** constitue sur ce thème une mise au point originale et très complète qui inclut des développements récents bien au delà de 1980. Il est bien sûr difficile d'identifier avec certitude le premier utilisateur d'une méthode qui repose sur des principes aussi simples (elle peut avoir été utilisée sans que cela donne lieu à une publication officielle), mais la variété des sources et la diversité des versions de l'algorithme est étonnante. Bien qu'elle ne fasse appel qu'à un formalisme limité et que son efficacité soit dans une large mesure attestée par les seuls résultats expérimentaux, la méthode d'*agrégation autour de centres mobiles* est probablement la technique de partitionnement la mieux adaptée actuellement aux vastes recueils de données ainsi que la plus utilisée pour ce type d'application. L'algorithme peut être imputé principalement à Forgy (1965), bien que de nombreux travaux (parfois antérieurs: Thorndike, 1953, est souvent cité, mais, comme le fait

remarquer H. H. Bock dans sa contribution, le lien avec la vrai méthode k-means n'est pas très clair), le plus souvent postérieurs (MacQueen, 1967, Diday, 1971) aient été menés parallèlement et indépendamment pour introduire des variantes ou des généralisations.

3.2 Classification hiérarchique, autres méthodes

Un premier historique sur la classification hiérarchique est dans l'article précité de Cormack (1971), mais une synthèse de travaux plus récents est celle de Gordon (1987).

Comme la classification autour de centres mobiles, la classification hiérarchique est souvent utilisée comme complément des méthodes d'analyses en axes principaux. Elle est donc abordée par plusieurs auteurs de ce numéro, notamment J. Gower qui retrace en tant que témoin de première ligne (section « Taxonomy » de son article) le premières tentatives et approches de Sneath, Williams, Lambert, Lance.

Mentionnons un point d'histoire intéressant : Il a été montré (Gower et Ross, 1969) que la classification hiérarchique suivant le critère du saut minimum (single linkage) équivalait au calcul de l'arbre de longueur minimale (minimum spanning tree) calculé sur le graphe complet valué par les distances entre objets. L'algorithme de Kruskal (1956) pour calculer cet arbre est le plus souvent cité. Mais l'algorithme de Florek *et al.*(1951), plus ancien et beaucoup plus performant, pouvant donner lieu à des calculs manuels, a donné lieu à une véritable école d'analyse des données en Pologne (la *Wroclaw taxonomy*) avant même la disponibilité des moyens de calcul actuels. On consultera Graham et Hell (1985) pour une histoire (passionnante) de l'algorithme de recherche de l'arbre de longueur minimum dont la première version actuellement recensée remonte en fait à Borůvka (1926).

D'autres critères d'agrégation que le saut minimum donnent éventuellement des résultats parfois plus fiables, évitant les « effets de chaîne » (cf. Ward, 1963 ; Wishart, 1969). Dans certaines conditions, ces algorithmes peuvent être grandement accélérés en utilisant le concept de voisins réciproques introduit par McQuitty (1966).

La contribution de **Boris Mirkin** et **Ilya Muchnik [Some topics of current interest in clustering: Russian approaches 1960-1985]** a le grand mérite de rendre accessible des références parfois ignorées des non russophones (les références sont parfois inconnues, mais les auteurs sont souvent connus!). Les thèmes traités (avec modestie et humour, au niveau de certaines réflexions des auteurs) sont les comparaisons de clusters, les études de consensus entre classifications, le "one cluster clustering", le "bi-clustering", les classifications sur les graphes.

4 Validité et portée des résultats

L'étude de la validité des résultats *des analyses en axes principaux* a donné lieu à des recherches nombreuses, mais prennent depuis 1980 des directions différentes avec les « *computer intensive methods* ». Elles sont évoquées par J. Gower dans sa contribution, mais peu présentes dans les diverses contributions de ce numéro. Il s'agit pourtant de recherches en tous points remarquables du point de vue théorique.

4.1 Stabilité numérique des axes principaux :

D'un point de vue purement numérique, Escofier et Le Roux (1972), Escofier (1979) ont traité de la stabilité des facteurs en analyse en axes principaux (analyse en composantes principales

et en analyse des correspondances). Ces auteurs étudient les variations maximales des facteurs et des valeurs propres lorsque l'on apporte des modifications bien déterminées aux données : suppression ou ajout d'éléments au tableau de données, influence du regroupement de plusieurs éléments ou de petites modifications des valeurs du tableau, influence du choix de la métrique et de la pondération. Leurs résultats se fondent sur les travaux de Davis et Kahan (1970). Les sous-espaces correspondant au haut du spectre sont les plus stables vis-à-vis des éventuelles perturbations de la matrice à diagonaliser (cf. les ouvrages fondamentaux de Wilkinson, 1965, et de Kato, 1966).

4.2 Inférence statistique et axes principaux

La loi de Wishart, établie par Fisher (1915) dans le cas de 2 variables, puis de façon plus générale par Wishart (1928), concerne la distribution d'une *matrice des covariances* empiriques. La densité de probabilité des *valeurs propres* issues d'une matrice de Wishart a été explicitée simultanément par Fisher (1939), Girshick (1939), Hsu (1939) et Roy (1939), bel exemple de découvertes simultanées et quasi-indépendantes. L'intégration de cette densité complexe a donné lieu à plusieurs publications ; parmi les principales, celles de Pillai (1965), Sugiyama (1966), Krishnaiah et Chang (1971), qui s'inspirent des travaux de Mehta (1960).

Dans sa publication donnant l'expression de la densité des valeurs propres d'une matrice de Wishart, Girshick (1939) calcule les variances et covariances asymptotiques (quand le nombre d'observations n tend vers l'infini) des éléments propres de la matrice des covariances expérimentales S, ceci dans le cas où la matrice des covariances théoriques Σ a toutes ses valeurs propres distinctes. Il donne également les variances et covariances théoriques des valeurs propres de la matrice des corrélations expérimentales, lorsque la matrice de corrélation théorique R a également toutes ses valeurs propres distinctes.

Bartlett (1950) propose une méthode pour tester l'égalité de $p-q$ valeurs propres des matrices Σ ou R. Lawley (1956) approfondit le cas des $p-q$ plus petites valeurs propres de Σ. Après ses travaux de pionniers sur les lois asymptotiques en 1951, Anderson (1963) a généralisé les résultats de Lawley en déterminant les lois limites des valeurs propres sans supposer que les valeurs théoriques correspondantes sont distinctes. Les intervalles de confiance d'Anderson sont encore très utilisés par les praticiens de l'analyse en composantes principales.

4.3 Le rééchantillonnage

Le bootstrap n'est rien d'autre qu'une technique de simulation particulière, fondée sur la distribution empirique de l'échantillon de base. Cette technique, introduite par Efron (1979), consiste sous sa forme la plus simple, pour estimer la confiance que l'on doit accorder à l'estimation θ^* d'un paramètre inconnu θ, à simuler m échantillons de même taille n que l'échantillon initial. Si l'on excepte les travaux de Gifi (1981) (voir la contribution de W. Heiser) qui concernent plus spécifiquement l'analyse des correspondances (le principe du bootstrap est en fait sensiblement différent selon les méthodes factorielles) les premiers travaux d'application du bootstrap à la validité des résultats en analyse en composantes principales sont ceux de Diaconis et Efron (1983), et donc postérieurs à la date limite que nous nous sommes fixée.

5. Thèmes connexes

Les contributions présentes dans ce numéro, comme ce rappel introductif n'ont pas l'ambition de couvrir tous les champs d'une discipline qui, même avant 1980, s'est développé dans de nombreuses directions. Des lacunes ou des analyses trop rapides subsistent, ce qui était inévitable. Citons par exemple les méthodes exploratoires dites de *Projection Pursuit* (Friedman et Tukey, 1974) qui connurent par la suite d'importantes extensions. Autre article de pionnier aux conséquences importantes à propos de l'analyse en composantes principales et des méthodes voisines : Wold (1966) (cité cependant par deux contributeurs à ce numéro : J. Gower et A. de Falguerolles). D'autres méthodes de description qui ne rentrent pas dans les deux grandes familles étudiées ici (axes principaux et classification) ne seront que brièvement évoquées, comme les méthodes purement graphiques, dévolues à la représentation de tableaux de petites dimensions, les méthodes de sériation, et surtout les méthodes de *multidimensional scaling* (Shepard, 1974, Kruskal et Wish, 1978) évoquées plusieurs fois par J. Gower dans sa contribution, mais non étudiées dans le détail de leur histoire.

Parmi les méthodes exclusivement visuelles, citons, surtout pour leur importance historique, les méthodes purement graphiques préconisées par Bertin (1967), la méthode des visages de Chernoff (1973), pour laquelle chaque visage correspond à un individu (ou une observation) et chaque trait du visage à une variable; la méthode des courbes d'Andrews (1972), où les différents paramètres des courbes sont les valeurs des variables; la méthode des constellations de Wakimoto et Taguri (1978). Les méthodes de sériations visent à faire apparaître des structures particulières de tableaux par simple reclassement de lignes et de colonnes. Pour des exposés de synthèse sur ce sujet, cf. par exemple Arabie (1978). Toutes ces méthodes graphiques interviennent souvent dans des contextes particuliers d'application et sont moins adaptées aux traitements des grands tableaux de données.

Bibliographie

Anderberg M.R. (1973): *Cluster Analysis for Applications*. Academic Press, New York.

Anderson T. W. (1951): The asymptotic distribution of certain characteristic roots and vectors. *Proc. of the 2nd Berkeley Symp. on Math. Statist. and Prob*., p 103-130, Univ. of California Press.

Anderson T. W., Rubin H. (1956): Statistical Inference in factor analysis. *Proc. of the 3rd Berkeley Symp. on Math. Statist.,* 5, p 111-150.

Anderson T.W. (1958): *An Introduction to Multivariate Statistical Analysis* (Second edition : 1984). J. Wiley, New York.

Anderson T. W. (1963): Asymptotic theory for principal component analysis. *Ann. Math. Statist.,* 34, p 122-148.

Andrews D. F. (1972): Plots of High-dimensional data. *Biometrics*, 28, p 125-136.

Arabie P. (1978): Constructing blockmodels : how and why. *J. of Math. Psychology,* 17, (1), p 21-63.

Bartlett M.S. (1950): Tests of significance in factor analysis. *British J. Psych. (Stat. Section),* 3, p 77-85.

Bastin C., Benzécri J.-P., Bourgarit C., Cazes P. (1980) : *Pratique de l'analyse des données*. Dunod, Paris.

Beltrami E. (1873): Sulle funzioni bilineari. *Giorn. Math. Battaglin.* 11, p 98-106.

Benzécri J.-P. (1969): Statistical analysis as a tool to make patterns emerge from clouds. In : *Methodology of Pattern Recognition* (S.Watanabe, Ed.) Academic Press, 35-74.

Benzécri J.-P. (1973): *L'Analyse des Données*. Tome 1: *La Taxinomie*. Tome 2: *L'Analyse des Correspondances* (2de. éd. 1976). Dunod, Paris.

Benzécri J.-P. (1982): *Histoire et préhistoire de l'analyse des données.* Dunod, Paris.

Benzécri J.-P. (1992): *Correspondence Analysis Handbook.* (Translation by T. K. Gopalan). Marcel Dekker, New York.

Bertin J. (1967) : *La sémiologie graphique.* Gauthier-Villars, Paris. (*The semiology of Graphics*. University of Wisconsin Press, 1983).

Bock H. H. (1974): *Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung und Strukturierung van Daten (Cluster Analysis)*. Vandenhoeck & Ruprecht, Götingen.

Borůvka O. (1926) : O jistém problému minimálním. *Práce Mor. Přírodověd. Spol. v Brně (Acta Societ. Scient. Natur.Moravicae)*, 3:37-58, 1926.

Burt C. (1950): The factorial analysis of qualitative data. *British J. of Statist. Psychol*. 3, 3, 166-185.

Carroll J. D. (1968): Generalization of canonical correlation to three or more sets of variables. *Proc. Amer. Psychological Assoc.* 227-228.

Chernoff H. (1973): The use of faces to represent points in *k*-dimensional space graphically. *J. Amer. Statist. Assoc.*, 68, 361-368.

Cliff N. (1966): Orthogonal rotation to congruence. *Psychometrika*, 31, 33-42.

Cormack R.M. (1971): A review of classification. *J. of Royal Statist. Society*, Serie A, 134, Part. 3, 321-367.

Cox D. R. (1977): The role of significance tests. *Scandinavian Journal of Statist*., 4, 49-70.

Davis C., Kahan W. M. (1970): The rotation of eigenvectors by a perturbation. *Journal of SIAM (Numerical Analysis)*, 7, 1-46.

De Leeuuw (1983): On the Prehistory of Correspondence Analysis. Statistica Neerlandica, vol 17, n° 4, 161-164.

Diaconis P., Efron B. (1983): Computer intensive methods in statistics. *Scientific American*, 248, (May), 116-130.

Diday E. (1971): La méthode des nuées dynamiques. *Revue Statist. Appl.* 19, n° 2, 19-34.

Eckart C., Young G. (1936): The approximation of one matrix by another of lower rank. *Psychometrika*, l, 211-218.

Efron B. (1979): Bootstrap methods : another look at the Jackknife. *Ann. Statist.*, 7, 1-26.

Escofier B. (1979 a): *Stabilité et approximation en analyse factorielle.* Thèse d'Etat, Université Pierre et Marie Curie, Paris.

Escofier B. [Cordier B.] (1965): *l'Analyse des correspondances.* Thèse, Faculté des Sciences de Rennes ; publiée en 1969 dans les *Cahiers du Bureau Universitaire de Recherche Opérationnelle*, n°13.

Escofier B., Le Roux B. (1972): Etude de trois problèmes de stabilité en analyse factorielle. *Publication de l'Institut Statistique de l'Université de Paris*, 11, 1-48

Fisher R. A. (1915): Frequency distribution of the value of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10, 507-521.

Fisher R.A. (1939): The sampling distribution of some statistics obtained from non linear equations. *Ann. Eugen.*, 7, 179-188.

Fisher R.A. (1940): The precision of discriminant functions. *Ann. Eugen*., 10, 422-429.

Florek K, Lukaszewicz J, Perkal J, Steinhaus H, Zubrzycki (1951): Sur la liaison et la division des points d'un ensemble fini. *Colloq. Math.*, 2, 282-285.

Forgy E. W. (1965): Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometric Society Meetings*, Riverside, California (Abstract in : *Biometrics* 21, 3, 768).

Friedman J. H., and Tukey J.W. (1974): A Projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, Ser. C, 23, 881-889.

Gabriel K.R. (1971): The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58, 3, 453-467.

Garnett J.-C. (1919): General ability, cleverness and purpose. *British J. of Psych.,* 9, 345-366.

Gifi A. (1981): *Non Linear Multivariate Analysis*, Department of Data theory, University of Leiden. Cf also: Gifi A. (1990): *Non Linear Multivariate Analysis*, Wiley, Chichester.

Girshick M.A. (1939): On the sampling theory of roots of determinantal equations. *Ann. Math . Statist.*, 1, 10, 203-224.

Gordon A.D. (1987): A review of hierarchical classification, *J.R.Statist.Soc.*, A, 150, Part 2, 119-137.

Gower J. C. (1966): Some distance properties of latent and vector methods used in multivariate analysis. *Biometrika*, 53, 325-328.

Gower J. C. (1975): Generalized Procrustes Analysis. *Psychometrika*, 40, (1), 33-51.

Gower J. C., Ross G. (1969): Minimum spanning trees and single linkage cluster analysis. *Appl. Statistics*, 18, 54-64.

Graham R. L. and Hell P. (1985) – On the history of the minimum spanning tree problem. *Ann. Hist. Comput*. 7, 43-57.

Guttman L. (1941): The quantification of a class of attributes: a theory and method of a scale construction. In : *The prediction of personal adjustment* (Horst{ XE "Horst" } P., ed.) 251 -264, SSCR New York.

Hand D. (1994): Review of the book: Correspondence Analysis Handbook (J.P. Benzécri). *Journal of Classification*, vol. 11, n°2, 289-290.

Harman H.H. (1967): *Modern Factor Analysis* (2nd ed.). Chicago University Press, Chicago.

Hartigan J. A. (1975) *Clustering Algorithms*. J. Wiley, New York.

Hayashi C.(1956) : Theory and examples of quantification. (II) *Proc. of the Institute of Statist. Math.* 4 (2), 19-30.

Hill M.O. (1974): Correspondence analysis: a neglected multivariate method. *Appl. Statist.* 3, 340-354.

Hirschfeld H.D. (1935): A Connection between correlation and contingency. *Proc. Camb. Phil. Soc.* 31, 520-524.

Horst P. (1961): Relation among m sets of measures. *Psychometrika,* 26, 129-149.

Horst P. (1965): *Factor Analysis of Data Matrices*. Holt, Rinehart, Winston, New York.

Hotelling H. (1933): Analysis of a complex of statistical variables into principal components. *J. Educ. Psy.* 24, 417-441, 498-520.

Hsu P. L. (1939): On the distribution of the roots of certain determinantal equations. *Ann. Eugen.* 9, 250-258.

Jardine N. Sibson R. (1971) – *Principle of Mathematical Taxonomy*. Wiley, New York.

Jordan C. (1874): Mémoire sur les formes bilinéaires. *J. Math. Pures et Appliquées*. 19, 35-54.

Kato T. (1966): *Pertubation Theory for Linear Operators.* Springer, New York.

Kendall M. G. (1966): Discrimination and classification. In : *Proc. Symp. Mult. Analysis*. Dayton, Ohio, (Krishnaiah P. R. (ed.), Academic Press, New York, 165-185.

Kendall M.G., Stuart A. (1961): *The Advanced Theory of Statistics*. Charles Grifin, London.

Kettenring R. J. (1971): Canonical analysis of several sets of variables. *Biometrika,* 58, (3), 433-450.

Krishnaiah P.R., Chang T. C. (1971): On the exact distribution of the extreme roots of the Wishart and MANOVA matrix. *J. of Multivariate Anal.,* 1, (1), 108-116.

Kroonenberg P. (1983): *Three-Mode Principal Component Analysis.* DSWO Press, Leiden.

Kruskal J. B. (1956): On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Amer. Math. Soc.* ,7, 48-50.

Kruskal J. B., Wish M. (1978): *Multidimensional Scaling*. Sage University Paper, 11, Sage, Beverly Hills.

Lawley D. N. (1956): Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika,* 43, 128-136.

Lawley D. N., Maxwell A. E. (1963): *Factor Analysis as a Statistical Method.* Methuen, London.

Lerman I. C. (1970): *Les Bases de la Classification Automatique.* Gauthier-Villars, Paris.

MacQueen J. B. (1967): Some methods for classification and analysis of multivariate observations. *Proc. Symp. Math. Statist. and Probability (5th),* Berkeley, 1, 281-297, Univ. of Calif. Press, Berkeley.

McQuitty L.L. (1966): Single and multiple classification by reciprocal pairs and rank order type. *Educational Psychology Measurements*. 26, 253-265.

Mehta M.L. (1960): On the statistica1 properties of the leve1 spacing in nuclear spectra. *Nucl. Phys.* 18, 395-419.

Mulaik S. A. (1972): *The Foundation of Factor Analysis.* McGraw Hill{ XE "Hill" }, New York.

Nakache J.P. (1973): Influence du codage des données en analyse factorielle des correspondances. Etude d'un exemple pratique médical. *Revue Statist. Appl*., 21, (2).

Nishisato S.(1980): *Analysis of Categorical Data. Dual Scaling and its Application*. Univ. of Toronto Press.

Pagès J.-P., Escoufier Y., Cazes P. (1976): Opérateurs et analyse de tableaux à plus de deux dimensions. *Cahiers du BURO,* ISUP, Paris, 61-89

Pearson K. (1901): On lines and planes of closest fit to systems of points in space. *Phil. Mag.* 2, n°ll, 559-572.

Pillai K.C.S. (1965): On the distribution of the largest root of a matrix in multivariate analysis. *Biometrika,* 52, 405-414.

Rao C.R. (1964): The use and interpretation of principal component analysis in applied research. *Sankhya* serie A, 26, 329-357.

Richardson M., Kuder G. F. (1933): Making a rating scale that measures. *Personnel Journal*., 12, 71-75.

Robert P., Escoufier Y. (1976): A unifying tool for linear multivariate methods : the Rv coefficient. *Applied Statistics,* 25, (3), 257-265.

Roy S.N. (1939): *p*-Statistics or some generalisations of analysis of variance appropriate to multivariate problems. *Sankhya,* 4, 381-396.

Schönemann P. H. (1968): On two-sided orthogonal procrustes problems. *Psychometrika,* 33, 19-33.

Schönemann P. H., Carroll R. M. (1970): Fitting one matrix to another under choice of a central dilation an a rigid motion. *Psychometrika*, 35, 245-255.

Shepard R. N. (1974): Representation of structure in similarity data : problems and prospects. *Psychometrika*, 39, (4), 373-421.

Sneath P. H. A. (1957): The Application of computers to taxonomy. *J. General Microbiology*, 17, 201-226.

Sokal R. R., Sneath P. H. A. (1963): *Principles of Numerical Taxonomy*, Freeman and co., San-Francisco{ XE "Francisco" }.

Spearman C. (1904): General intelligence, objectively determined and measured. *Amer. Journal of Psychology,* 15, 201-293.

Sugiyama T. (1966): On the distribution of the largest latent root and the corresponding latent vector for principal component analysis. *Ann. Math. Statist.* 37, 995-1001.

Sylvester J.J. (1889): *Messenger of Mathematics* (cité par Eckart, Young, 1939). 19, n°42.

Tenenhaus M., Young F. W. (1985): An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 91-119.

Thorndike R.L. (1953): Who belongs in the family. *Psychometrika*, 18, 267-276.

Thurstone L. L. (1947): *Multiple Factor Analysis.* The Univ. of Chicago Press, Chicago.

Tucker L. R. (1958): An inter-battery method of factor analysis. *Psychometrika,* 23, (2).

Tucker L. R. (1964): The extension of factor analysis to three-dimensional matrices. In : *Contribution to Mathematical Psychology*, Harris C. W. (ed.), Univ. of Wisconsin Press, Madison, 109-127.

Tucker L. R. (1966): Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31, 279-311.

Tukey J. W. (1977): *Exploratory Data Analysis*. Addison Wesley, Reading, Mass.

Wakimoto K., Taguri M. (1978): Constellation graphical methods for representing multidimensional data. *Ann. of the Inst. of Statist. Math.*, 30, (1), 97-104.

Ward J.H. (1963): Hierarchical grouping to optimize an objective function. *J. of Amer. Statist. Assoc.*, 58, 236-244.

Wilkinson J. H. (1965): *The algebraic Eigenvalue Problem*. Clarendon Press, Oxford.

Wishart D. (1969): Mode analysis : a generalization of nearest neighbour which reduces chaining effects. *Numerical Taxonomy* (A.J. Cole ed.) 282-311, Academic Press, London.

Wishart J. (1928): The generalized product-moment distribution in samples from a normal multivariate population. *Biometrika*, 20A, 32-43.

Wold H. (1966): Estimation of principal components and related models by iterative least squares, in *Multivariate analysis*, Krishnaiah{ XE "Krishnaiah" } P.R. (Ed.), Academic Press, New York, pp391-420.