LES CAHIERS DE L'ANALYSE DES DONNÉES

J. P. BENZÉCRI

Analyse des données en physique corpusculaire

Les cahiers de l'analyse des données, tome 3, n° 1 (1978), p. 79-94

<http://www.numdam.org/item?id=CAD_1978__3_1_79_0>

Les Cahiers de l'Analyse des Données
Vol III 1978 n° 1 p 79 94

ANALYSE DES DONNÉES EN PHYSIQUE CORPUSCULAIRE

III. – Les méthodes multidimensionnelles

[PHYS. COR.]

par J. P. Benzécri (1)

Après avoir rappelé dans un premier article (Vol II n° 3) le cadre théorique auquel il s'impose de rapporter les données recueillies en physique des hautes énergies, on a montré l'efficacité des techniques uni- ou bidimensionnelles appliquées couramment (Vol II ; n° 4). Cependant certains physiciens croient devoir employer de plus en plus l'analyse multidimensionnelle. Leurs méthodes souvent originales parfois conçues sans la collaboration des écoles statistiques, devraient inciter le spécialiste à expérimenter de nouveaux algorithmes, notamment en classification et pour l'ajustement et l'estimation des lois multidimensionnelles. On s'est donc appliqué à exposer ces méthodes sans faire à la physique des références difficiles à suivre. Un dernier § purement statistique (§ 4.5) confronte estimation robuste multidimensionnelle et analyse des données. Avec le présent article se termine l'exposé critique des recherches antérieures à 1976. Nous désirons poursuivre cette série en présentant des travaux nouveaux (P. Lutz ; D. Maïti) portant sur de très grands ensembles de données de physique corpusculaire ; et appliquant certaines des suggestions déjà faites ici.

4.3 *Partition d'un ensemble d'événements en classes* : Partager un échantillon d'événements d'un même canal en sous-échantillons rapportés à un ensemble C de sous-canaux est un problème de physique auquel il semble naturel d'appliquer les techniques statistiques de la classification automatique. Un certain nombre de physiciens pensent avec W. Kittel (cf Prague 1974) , que ces techniques seront indispensables pour découvrir des structures encore inconnues d'après les grands échantillons d'événements comportant de nombreuses particules émergentes (e.g. dix ou plus); échantillons que fourniraient les expériences à très haute énergie réalisables avec les accélérateurs qui entrent présentement en service. Jusqu'ici il apparaît que la classification automatique a été presque exclusivement appliquée à des échantillons d'événements à 3 ou 4 particules émergentes, dans des canaux déjà bien connus ; il ne s'agit que d'essais par lesquels les physiciens s'accoutument à une technique nouvelle en critiquant les résultats qu'elle fournit d'après des connaissances solides qu'ils ont acquises par ailleurs.

En bref, selon W. Kittel (1974) deux conclusions se dégagent quant à la physique :

a) un mécanisme physique unique peut se trouver scindé en deux ou plusieurs classes ;

b) une même classe d'événements peut recevoir des contributions de deux ou plusieurs mécanismes physiques.

Il n'y a pas de correspondance biunivoque entre classes et mécanismes physiques (sous-canaux etc.). A la vérité ce que nous savons de la structure d'un canal (cf § 4.1) ne laissait pas espérer une telle correspondance. Au contraire il est intéressant de découvrir le morcellement (cf a) d'un sous-canal en plusieurs classes que séparent les lignes

(*) Suite de l'article paru sous le même titre dans les Cahiers, Vol II n° 4.
(1) Professeur de Statistique. Université Pierre et Marie Curie, Paris.

nodales de son amplitude (lignes sur V où A(v) s'annule) ; et aussi disposer de certaines classes (cf b) où les interférences entre sous-canaux jouent à plein.

Il ne nous appartient pas d'assombrir les vues assez optimistes de W. Kittel. Dnas ce § nous considérerons quelques principes d'analyses des données déjà mis en oeuvre pour l'étude des réactions à haute énergie en complétant et systématisant un arsenal de méthodes qui semble s'être constitué jusqu'ici au hasard des rencontres entre physiciens et statisticiens. Il s'agit principalement de techniques de classification. On peut s'étonner que l'analyse factorielle d'un nuage d'événements n'ait pas encore été tentée pour en séparer les bandes et les branches. Au § 4.3.1, on verra que l'analyse factorielle typologique (Y. Ok, thèse 3° cycle, Paris 1975) semble proche de certaines démarches des physiciens ; et peut-être l'analyse factorielle déjà utilisée pour la détection des trajectoires (§ 4.4.1) servira-t-elle-aussi un jour à la découverte et à la description des sous-canaux.

4.3.1 L'agrégation autour des noyaux variables : Rappelons brièvement le principe d'une famille d'algorithmes qui sont croyons-nous compris et utilisés au mieux dans les travaux de E. Diday et de ses collaborateurs Y. Ok, A. Schroeder (Madame Huet).

Exposons d'abord sommairement l'algorithme des nuées dynamiques de E. Diday. Soit I un ensemble d'individus (points d'une variété ou espace V) ; supposons donné dans V un ensemble C de points c appelés centres; en rattachant chaque individu i au centre c dont il est le plus proche on obtient une partition de I en classes I_c chacune agrégée autour d'un centre $c \in C$. Maintenant il est possible de calculer suivant tel ou tel principe à fixer (e.g. comme centre de gravité) un centre $\mathcal{C}(I_c)$ pour chacune des classes obtenues. Ces centres peuvent être substitués aux centres c pris d'abord :

$$C^\circ = C \to \{ \mathcal{C}(I_c) \mid c \in C \} = C^1 ;$$

et autour des centres du système C^1 on peut obtenir comme précédemment une partition de I en un système de classes I^1_c chacune centrée sur un $\mathcal{C}(I_c) \in C^1$. Le processus peut être poursuivi itérativement, jusqu'à ce que la partition ne se modifie plus, ou se modifie peu.

Plus généralement, au lieu de centres ponctuels c, on peut utiliser des noyaux N_c : l'algorithme d'agrégation autour de noyaux variables requiert seulement deux constructions. 1°) Etant donné une partie I_c de I, déterminer un noyau $\mathcal{N}(I_c)$ qui en sera le centre. 2°) Etant donné un système de noyaux $\{N_c \mid c \in C\}$, partager l'ensemble I, en classes I_c en affectant chaque individu i à celui des noyaux N_c dont il est, en un certain sens le plus proche.

Voici deux exemples de noyaux. Un noyau de loi normale. (A. Schroeder thèse ; Paris 3° cycle 1974 est donné par un centre g_c, une matrice des variances covariances σ_c, et une masse totale m_c : à un tel noyau est associée une densité continue. Il est facile d'associer un noyau de loi normale à toute partie finie I_c. Et réciproquement étant donné un système de noyaux, un individu i sera rattaché à celui des noyaux dont la densité en i est la plus forte. Il est également possible de prendre pour noyaux des sous-variétés linéaires : droites, plan etc... (Y. Ok; thèse Paris 3° cycle 1975). Evidemment un individu i sera rattaché à la sous-variété linéaire dont il est le plus proche. Quant au noyau d'une classe I_c on le définit e.g. par la condition des moindres carrés : si le noyau doit être une droite, on prendra le premier axe principal

d'inertie du sous-nuage I_c ; si ce doit être un plan on prendra le plan engendré par les deux premiers axes principaux d'inertie ; etc. Déterminer le noyau centre d'une classe requiert donc qu'on fasse une analyse factorielle de cette classe : c'est pourquoi Y. Ok parle d'analyse factorielle typologique.

Nous avons vu (§ 4.2.4 *in fine*) que l'agrégation autour de centres variables fait partie intégrante du NAMELESS project de E. Pagiola. Le même principe est à l'oeuvre dans les travaux présentés à Genève par divers auteurs ; notamment par Ch. de la Vaissière *(Application of the CERN interactive cluster analysis to multibody $\bar{p}$ p interactions)*. Toutefois il semble qu'à chaque itération, pour définir les noyaux centres N_c des classes qu'ils viennent de construire, ces auteurs ne laissent jamais l'ordinateur opérer seul suivant un critère mathématique préétabli ; l'opérateur humain intervient pour fixer des bornes d'après des histogrammes (et c'est pourquoi Ch. de la V. parle d'analyse interactive). Au contraire H. Schiller (cf H. S. & W.D. Nowak : *A heuristic cluster algorithm for the analysis of many particle final states* ; article destiné *aux Computer Physics Communications)* utilise en substance l'algorithme d'agrégation autour de noyaux de loi normale qu'A. Schroeder a programmé avant lui ; mais il complète cet algorithme par des essais de subdivision et combinaison de classes dont nous parlerons au § 4.3.3. De même dans la *Valley seeking technique* dont les applications furent présentées par K. Lanius et P. Lutz, l'analyse est entièrement automatique. Mais pour expliquer cette *technique*, il nous faut rappeler encore un autre principe d'algorithme (principe déjà classique, croyons-nous, comme celui des noyaux variables) qui sert ici à agréger les individus autour des noyaux.

4.3.2 L'algorithme local d'affectation majoritaire et de régression :

En terme imagé, on parle aussi de discrimination ou de régression par *boule*. Expliquons succintement ces algorithmes. Soit I un ensemble d'individus i dont chacun est affecté à une classe c(i) ; soit s un individu supplémentaire (éventuellement $s \in I$) ; considérons dans la variété V ambiante une boule B(s,R) de centre s, dont le rayon R est choisi en sorte que B(s,R) contienne e.g. 10 (ou 20) individus de I : on décid'affecter l'individu s à celle des classes c qui compte dans B(s,R) le plus d'individus i : c'est l'affectation majoritaire locale, ou par *boule*. Pour faire une régression par boule, supposons définie sur I une certaine fonction x ; la valeur de x(s) sera calculée en prenant la moyenne des valeurs x(i) pour les individus i contenus dans B(s,R) (*).

La *Valley seeking technique* suggérée par Koontz & Fukunaga (*IEEE Transactions on Computers* C Rl, (1972), 2) et expérimentée en physique des hautes énergies par Böttcher, Kosta, Lanius, Roloff & Schiller sur la réaction $\pi^+ p \to p\pi^+\pi^+\pi^-$ (à 8 GeV), combine le principe des centres variables à celui de l'affectation majoritaire. Soit I un ensemble d'individus, répartis (au départ aléatoirement) en des classes I_c. (e.g. trois classes I_1 , I_2 , I_3). Pour attribuer à chaque individu, i, une nouvelle affectation K. & F. considèrent la boule B(i,R), et ils réaffectent i à celle des classes I_c qui compte dans cette boule le plus d'individus ; et ainsi de suite itérativement. On peut dire qu'ici les noyaux sont des ensembles des parties de I : le noyau central (I_c) d'une partie I_c de I n'est autre que la classe I_c elle-même ; on dit que i est d'autant plus proche de I_c qu'il y a plus d'individus de I_c dans la boule b(i,R).

Quant aux mérites de la technique de K. & F., les avis semblent partagés : P. Lutz lui paraît peu favorable. Un détail critique, est que K. & F. proposent de choisir un même rayon R pour tous les centres i ,

(*) *Là où nous parlons de* boule, *les auteurs anglo-saxons parlent de* nearest neighbour, *ou h-ème voisin par ordre de proximité.*

en sorte qu'il faut de laborieux tâtonnements pour trouver un R satisfaisant. Dans l'exposé de l'algorithme de la boule, nous avons au contraire spécifié de choisir R pour chaque individu de telle sorte que B(i,R) contienne e.g. 10 ou 20 points de I : ce nombre (10 ou 20) est choisi pour s'affranchir des fluctuations d'échantillonnage sans élargir plus qu'il n'est indispensable le voisinage (boule) que l'on inspecte pour décider de i.

Cependant le lecteur s'interrogera sur le terme même de *Valley seeking technique*, technique de recherche des vallées. Pour K. & F. il s'agit d'aboutir à un système de classes bien séparées par des plages vides, des vallées. Mais, pour atteindre cet idéal, l'algorithme de K.&F. ne considère pas explicitement la densité du nuage I. Voyons maintenant comment celle-ci peut être prise en compte.

4.3.3 Densité du nuage des événements et distance dans l'espace des phases :

La densité d'une distribution de masse ν ne peut être définie que par rapport à une autre distribution de masse μ, qu'on peut appeler l'élément de volume de référence. De plus il faut que ν soit continue par rapport à μ. Dans le cas du nuage (ou échantillon) I des événements il s'impose de prendre pour mesure de référence l'élément de volume naturel dont est munie la sous-variété V des éléments permis dans l'espace des phases (cf § 3.2) ; mais de plus, puisque I est un ensemble fini de masses ponctuelles, non une distribution continue de masses, il faut de quelque manière procéder à un lissage. Pour évaluer la densité du nuage I au voisinage d'un point s (qui peut ou non appartenir à I) considérons comme au § 4.3.2 une boule B(s,R) de centre s contenant, e.g., 20 points du nuage I ; si vol(B(s,R)) désigne le volume naturel de cette boule dans V, on pourra prendre pour estimation de la densité en s le quotient $20/\mathrm{vol}(B(s,R))$. cette estimation ne peut être très précise ; car le nombre de points de I dans B(s,R) est sujet à des fluctuations d'échantillonnage dont la variance est $\sqrt{20}$: i.e. un autre échantillon I' (de même effectif que I) pourra avoir dans B(s,R) non 20 points mais 16 ou 24. De plus au lieu de l'expression analytique de l'élément de volume naturel on peut préférer engendrer par simulation (méthode dite de Monte-Carlo cf § 4.1.2) un échantillon U de densité uniforme ; dès lors la densité sera estimée comme le quotient du nombre des individus de I contenus dans B(s,R) par le nombre des individus de U contenus dans cette même boule ; et alors le numérateur et le dénominateur seront tous deux sujets à fluctuations. Mais une estimation imprécise de la densité suffit à reconnaître dans I les zones les plus denses ; et dans l'espace ambiant les plages presque vides : cette information qualitative peut aider à la classification automatique. Par exemple avant de soumettre I à la classification, on en éliminera 20% des individus ; choisis pour être ceux où la densité est la plus faible : ainsi se trouvera accru le contraste entre zones denses et plages vides. Si de plus le programme de classification fait usage de noyaux variables (§4.3.1) on partira non de noyaux tirés au hasard, mais de centres de densité maxima, etc...

Dans son rapport de 1974, W. Kittel, faisant référence à des idées de L. Van Hove décrit un algorithme de classification fondé sur une estimation de la densité analogue à celle que nous venons de proposer. Il propose alors d'agréger en classes le sous-ensemble I' de I, formé des individus i où la densité est supérieure à λ fois son maximum ; e.g. λ=0,1; selon W. K., λ est à fixer par tâtonnement en descendant à partir de la valeur 0,5, jusqu'à ce que des mécanismes physiques distincts se trouvent confondus dans une même classe. Au sein de I', les classes sont définies à partir de la notion de points contigus : deux points sont dits contigus si leurs voisinages se coupent ; les voisinages dont parle W.K. ne nous semblent pas définis de façon invariante, il faudrait ici un seuil de distance, ce qui requiert une métrique naturelle, dont nous reparlerons. Ceci posé on dira, en bref, que les classes de I' sont les parties maximales formées de points que l'on peut tous relier entre eux par des chaînes d'individus dont chacun est contigu à son prédécesseur et à son successeur.

Dans cet algorithme, (comme dans la *Valley seeking technique* exposée au § 4.3.2) un spécialiste de la classification automatique voit à l'oeuvre plusieurs idées. Nous avons déjà exposé l'estimation de la densité par une boule ; nous rappellerons au § 4.3.4 la procédure d'agrégation par le saut en classification ascendante. Quant au choix d'une distance naturelle, il est lié pour nous à la géométrie de l'espace des phases et à l'élément de volume naturel. Dans R^{4n} (produit de n espaces munis de la métrique hyperbolique), la sous-variété V des éléments permis est une sous-variété de type espace : toutefois l'élément de volume associé à la métrique induite par celle de R^{4n} n'est pas en général l'élément de volume invariant : il faut donc lui faire subir un changement d'échelle variable en chaque point, c'est-à-dire une transformation conforme, pour que métrique et élément de volume s'accordent. Pratiquement il n'est pas nécessaire en classification automatique, de calculer la distance d(i,i') par intégration d'un ds sur une ligne géodésique : il suffit de calculer d(i,i') par la formule hyperbolique usuelle de R^{4n}, mais en corrigeant par un changement d'échelle dont le coefficient est calculé d'après l'élément de volume naturel en i et i'. Par exemple supposons construit un échantillon U uniforme pour l'élément de volume naturel ; soit B(i,R), B(i',R') des boules (pour la métrique hyperbolique usuelle) contenant chacune 20 points de U : on pourra corriger la distance $\|i\,i'\|$ en la divisant par R + R', ou inf(R,R') etc... Les physiciens ont maintes fois fait remarquer (cf Kittel 1974) que les conclusions auxquelles aboutit l'analyse d'un ensemble I d'événements ne peuvent être que des artefacts si elles subsistent avec un échantillon fictif engendré sous l'hypothèse d'uniforme densité dans l'espace des phases (terme de référence dont on déplorera parfois cependant qu'il soit très éloigné des faits ; cf § 4.2.2 *in fine).*

L'idéal que nous poursuivons est d'incorporer dans la méthode statistique elle-même toutes les propriétés d'uniformité et d'invariance que la physique donne pour hypothèse nulle : c'est pourquoi ayant calculé la densité par rapport au volume naturel, nous voulons aussi que l'agrégation soit faite suivant une distance naturelle.

De plus les calculs de masse invariante (§ 3.3) jouent un rôle essentiel dans la découverte des sous-canaux (§ 3.5) on souhaite en tenir compte dans le calcul de la distance entre événements. Reprenons l'exemple de la figure 3-5 (§ 3.5) : deux événements rapportés au sous-canal.

$$K^+ + p \to K^\circ + N^* \to K^\circ \qquad \pi^+ + p,$$

peuvent être considérés comme proches si la somme $\{p\}_{\pi+} + \{p\}_p = \{p\}_{N^*}$ des quadrimoments des particules émergentes censées provenir du N^* intermédiaire, varie peu de l'autre : car alors les deux éléments diffèrent seulement quant à la désintégration du N^*, qui est un processus secondaire. Plus généralement si pour un événement produisant un ensemble J de particules émergentes on adopte le schéma de la figure 4-3, on posera :

$$\{p\}_{J1} = \Sigma\{\{p\}_{ej} \mid j \in J_1\} \; ; \; \{p\}_{J2} = \Sigma\{\{p\}_{ej'} \mid j' \in J_2\} \; ;$$

et pour comparer deux événements ainsi décomposés on comparera principalement leurs quadri-moments partiels $\{p\}_{J1}$ et $\{p\}_{J2}$. Le calcul de la distance entre deux événements se ferait donc en choisissant la décomposition $J = J_1 \cup J_2$ pour laquelle ils apparaissent plus proches . Sans fixer ici de formule, nous suggérons à l'analyse statistique multidimensionnelle de s'inspirer des programmes interactifs (cf e.g. § 4.2.4 *The NAMELESS Project)* où l'opérateur contrôle sans cesse la cohésion des classes d'après les histogrammes de masses invariantes.

Dans ce même § 4.3.3, nous parlerons encore du *Projection Pursuit algorithm* programmé par J. Friedman selon les vues de Tuckey (*)et présenté

(*) *cf e.g. J.H. Friedman & W. Tuckey : A projection pursuit algorithm for exploratory data analysis; in* IEEE Trans. Comp. *Vol. G 18 pp 401-409 (1969).*

à Genève par P. Lutz ; car il s'agit d'assigner au nuage I une densité et de rechercher une plage vide où placer un hyperplan séparateur ; et nous terminerons sur une idée analogue mise en oeuvre par H. Schiller. Soit $\rho(x)\,dx$ la loi d'une variable aléatoire unidimensionnelle ; σ la variance de cette loi ; l'intégrale $\sigma\int\rho(x)^2dx$ est une grandeur sans dimension (en bref si l'unité de longueur est divisée par 2, dx est multiplié par 2 ainsi que σ ; mais ρ est divisé par 2) qui présente l'intérêt d'être d'autant plus grande que la loi $\rho(x)\,dx$ est moins unimodale (e. g. présente deux maxima étroits séparés par une large place vide). Donc l'algorithme recherche dans l'espace ambiant au nuage I une forme linéaire (combinaison des coordonnées initiales) pour laquelle soit maxima l'intégrale $\sigma\int\rho^2\,dx$. En fait le nuage I n'étant qu'un système de masses ponctuelles, il faut pour calculer une densité ρ de la distribution d'une variable x_i sur I procéder à un lissage (cf le début de ce § : plus précisément, pour estimer $\int\rho(x)^2dx$, les auteurs calculent $(1/\text{Card I})\,\Sigma_i\,\rho(x_i)^2 \approx \int\rho(x)\rho(x)dx$). La forme linéaire optimale est atteinte par un algorithme usuel de recherche d'optimum : c'est pourquoi les auteurs parlent de poursuite. Enfin cette forme étant trouvée, elle doit faire apparaître par projection une division du nuage en classes : e.g. si la combinaison $x + 2y - z$ (des trois coordonnées de base x, y, z de R^3) est sur le nuage I fortement multimodale c'est que I projeté sur une droite parallèlement à la direction du plan $x + 2y - z = 0$ donne deux (ou plusieurs) ilôts séparés ; ou encore I peut être partagé en deux classes bien distinctes par un plan $x + 2y - z = \text{cte}$. Telle est l'essence du *Projection pursuit algorithm.*

Nous avons dit que H. Schiller & W.D. Nowak dans leur algorihme heuristique, mettent en oeuvre, après A. Schroeder, l'agrégation des événements autour de noyaux variables de loi normale. Mais de plus afin de renouveler les classes plus énergiquement que ne le peut faire l'affectation itérative des points à des centres variables, H.S. & W.D.N. prévoient à chaque itération la possibilité de scinder une classe en deux ou de fondre deux classes en une. En bref ces auteurs définissent une quantité critère qui est un indice de bimodalité pour la somme de deux lois normales. Si la somme des lois normales (noyaux) ajustées à deux classes a un indice de bimodalité faible, ils décident de fondre celles-ci en une. Et corrélativement, ils tentent de couper chaque classe q en deux par des hyperplans tirés au hasard (passant par le centre de gravité g_q) : il en résulte deux classes (q^+ et q^-) de part et d'autre de l'hyperplan : à q^+ et q^- sont ajustées des lois normales noyaux ; si la somme de celles-ci a un indice de bimodalité élevé, le programme décide d'adopter les deux nouvelles classes q^+ et q^- en remplacement de la classe unique préexistante q.

4.3.4 <u>*Classification arborescente*</u> : Lors même qu'on ne désire qu'une partition de I en un ensemble fini de classes, la recherche directe de cette partition peut n'être pas la voie la plus rapide. Construire une classification hiérarchique semble un objectif plus ambitieux et une tâche plus lourde, que de construire une simple partition : mais une classification hiérarchique offre un grand nombre de partitions, entre lesquelles on peut choisir en critiquant la validité des classes associées aux différents noeuds ; et dans le cadre de la classification hiérarchique, ce choix se fait plus efficacement que si l'on tâtonne en examinant des partitions obtenues successivement pour plusieurs valeurs des paramètres-seuils. Certes il semble fabuleux de tenter une classification ascendante hiérarchique sur plusieurs milliers d'individus (quelques centaines est la limite supérieure de ce que les ordinateurs de 1976 traitent en un temps acceptable en appliquant l'algorithme usuel de classification ascendante. Mais d'une part on peut passer e.g. de 3000 à 300 par une procédure d'agrégation rapide ; puis faire sur les 300 ilôts obtenus une classification ascendante hiérarchique. ou faire choix d'un

sous-échantillon de 300 points (choix au hasard ; peut-être parmi les points où la densité est assez élevée) ; soumettre cet échantillon à la classification ascendante hiérarchique ; puis y adjoindre en éléments supplémentaires les 2700 individus écartés d'abord. Contre de telles stratégies, perd de sa force une objection plusieurs fois énoncée par J. Friedman au colloque de Genève : "cette méthode est en N^3 (i. e. requiert $\approx N^3$ opérations élémentaires pour traiter un échantillon de N points) ; elle doit donc être écartée puisqu'il existe des méthodes en N^2, voire en N ". D'autre part on dispose depuis grâce aux travaux de M. Bruynooghe (ce cahier Vol III n° 1, pp 7 - 33) d'un algorithme de classification ascendante hiérarchique dont le taux de croissance bien loin d'être en N^3 est inférieur même à N^2.

Voilà pourquoi nous suggérons d'appliquer aux données de physique corpusculaire la classification ascendante hiérarchique. On sait que l'algorithme ascendant, peut utiliser plusieurs procédures d'agrégation avec diverses formules de distance, le format général restant le suivant : d'abord unir en une classe c les deux individus i et i' de I qui sont les plus proches ; puis former une classe c' en unissant deux termes aussi proches que possible (soit deux points i" et i"' ; soit la classe c et un individu i") etc. On devine que les variantes de la méthode dépendent des sens divers que l'on peut donner à la locution "plus proche", ou au mot "distance" particulièrement distance entre classes.

L'algorithme exposé par Kittel (1974) et discuté au § 4.3.3, suggère de prendre pour procédure, l'agrégation suivant le saut minimum (on mesure la distance entre deux classes c et c' par le saut minimum qu'il faut faire pour passer de l'une à l'autre ; i.e. d(c,c') est le minimum de d(i,i') pour $i \in c$ et $i' \in c'$) ; en corrigeant la formule de distance entre points comme il est proposé au § 4.3.3. Pour les applications les plus récentes de la classification hiérarchique on se reportera à la thèse de D. Maïti (3° cycle Paris 1977) et aux travaux ultérieurs.

Nous terminerons ce § sur la classification arborescente en signalant que fut présenté à Genève par J. Schotanus un essai de taxinomie polonaise (ou recherche d'arbre de longueur minima) sur un ensemble d'événements à trois particules émergentes. On sait que cette technique consiste à relier entre eux certains couples d'éléments de l'ensemble I, en sorte que d'une part les liens ainsi établis permettent de relier par une chaîne tout i à tout i' et d'autre part la somme de leurs longueurs soit minima. Il en résulte un arbre (au sens qu'a ce mot en théorie des graphes : arbre = graphe sans circuits) mais non une classification arborescente hiérarchique (dans l'arbre associé à une telle classification les noeuds sont des classes, les individus apparaissent en position terminale, ce n'est pas le cas en taxinomie polonaise, où noeuds et éléments terminaux de l'arbre sont des individus à classer) ; et il est souvent fort difficile de dessiner clairement l'arbre sans que les traits se coupent. On sait d'autre part (cf Cahiers Vol I n° 4, pp 441 sqq.) du graphe de la taxinomie polonaise le passage est aisé à une véritable classification arborescente hiérarchique : mais celle-ci se trouve satisfaire au critère d'agrégation suivant le saut minimum (*single-linkage clustering* des auteurs anglo-saxons) ; critère dont on sait qu'il est en butte à un fort effet de chaînage (constitution de classe en chaînes, ou filiformes réunissant des points très éloignés qu'il conviendrait de distinguer. J. Schotanus présentait des diagrammes de Dalitz ou des vues tridimensionnelles, où les points i étaient à leur place, les liens de l'arbre de longueur minima étant tracés en plus, évidemment avec de nombreuses intersections produisant une sorte de résille, à notre avis mal interprétable. Plus intéressant peut être la suggestion de considérer le chemin le plus long du graphe (où par "chemin" l'on entend ligne brisée formée d'une succession sans répétitions de liens, ou segments de l'arbre de longueur minima) : parfois ce chemin fournit une sorte de squelette du nuage I.

4.4 *Diversité des problèmes statistiques* : La classification automatique peut être utilisée, on l'a dit au § 4.3, pour partager un échantillon d'événements en classes et reconnaître des sous-canaux. D'autres méthodes de statistique multidimensionnelle sont susceptibles de contribuer au progrès de la physique corpusculaire : nous en donnons quelques exemples dans ce §. Et concluons après avoir cité W. Kittel, en invitant le géomètre à fonder sur la forme précise des nuages d'événements une recherche des lois analytiques exactes de la physique.

4.4.1 *Détection des trajectoires et analyse factorielle* : Les clichés issus de chambres à bulles (cf § 2.3) contiennent des informations parfois confuses, mais toujours redondantes, en ce sens qu'une trajectoire s'y matérialise par un chapelet continu de bulles. Cependant, notamment pour observer les réactions entre particules de deux faisceaux qui se rencontrent et non entre un faisceau et une cible dense, on saisit aujourd'hui des trajectoires définies soit par une suite discontinue de points (chambres à étincelles), soit par leurs impacts dans un petit nombre de plans (chambres à fils). Corrélativement, l'aptitude d'un homme à reconnaître les trajectoires est de moins en moins satisfaisante, et le rôle de l'ordinateur devient prédominant. Selon P. Zanella *(Machine Recognition of Patterns in Particle Physics ;* CERN ; DD/72/16 ; Mai 1972), en physique des chambres à bulles l'acquisition totalement automatique des trajectoires (en balayant les clichés par un spot etc.) est réalisable, mais elle n'est pas économique : la seule solution pratique est la coopération de l'opérateur qui lit le cliché avec la machine qui enregistre les coordonnées des points qui lui sont désignés. Dans les chambres à étincelles, comme le nombre des points (étincelles éclatant entre des conducteurs sous tension au passage d'une particule) est beaucoup moindre que celui des bulles (*), le traitement automatique complet devient possible et certaines expériences auraient été inconcevables sans ce mode de dépouillement. Avec les chambres à fils (dispositifs divers, localisant par une étincelle ou une impulsion électrique le passage d'un corpuscule au niveau de plans, tendus de fils parallèles) on quitte décidément le domaine où l'homme peut concurrencer la machine. De plus, la reconnaissance des trajectoires, bénéficie grandement des méthodes statistiques multidimensionnelles : c'est ce que nous verrons sur l'exemple du spectromètre R603 utilisé actuellement au CERN.

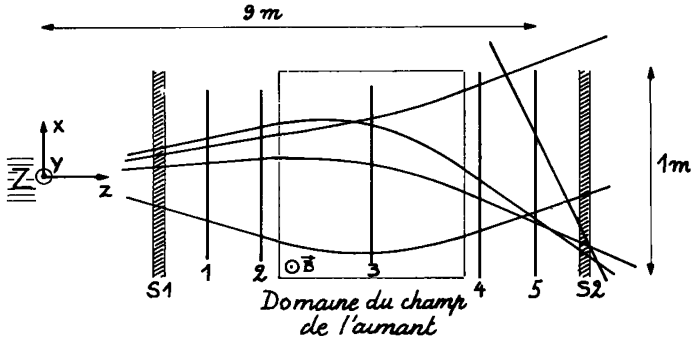

Figure 4-6 Schéma du spectromètre R-603 (d'après des documents reçus de J. Kubler). on a noté Z la zone d'intersection des faisceaux (origine des événements); S1 et S2 les scintillateurs; 1,2,3,4,5 les plans de détection.

(*) *Dans le spectromètre à étincelles Ω du CERN, une trajectoire est matérialisée par des chapelets d'étincelles espacées de 2,5 cm ; l'intervalle entre deux chapelets est de 20 cm ; au total le nombre des étincelles par trajectoire approche la centaine.*

Conçu pour étudier les réactions entre deux faisceaux de protons à très haute énergie ($\approx$ 30 GeV) dont les trajectoires, de sens opposés, se coupent sous un petit angle dans une zone Z, le spectromètre R603 (fig. 4-6) qui évite les deux faisceaux doit détecter dans cinq plans 1,2,3,4, 5, les impacts des particules émergentes. Entres les plans 2 et 4, règne un champ magnétique (à peu près constant) destiné à révéler la vitesse des particules par la courbure de leurs trajectoires (plus précisément on mesure le rapport mv/e de la quantité de mouvement mv à la charge e qui est toujours au signe près celle de l'électron). Afin de circonscrire l'observation et de s'affranchir autant que possible des impacts parasites de particules transversales (ne provenant pas de la zone E : e.g. rayons cosmiques), l'enregistrement des impacts est déclenché (en anglais : *triggered*) par la coïncidence de deux scintillateurs S1 et S2 : en bref si dans 1 ns (10^{-9} seconde) S1 et S2 ont été stimulés par la traversée de particules en haute énergie, les plans 1 à 5 sont mis sous tension (ces plans sont en fait des chambres de 7 cm d'épaisseur, comportant quatre plans tous quatre garnis de fils parallèles, chacun suivant sa direction ; et où l'ionisation provoquée par la particule provoque l'éclatement d'étincelles entre fils voisins de potentiels opposés) et les impacts sont enregistrés.

Cependant la connaissance des ces impacts, ne révèle pas immédiatement les trajectoires des particules ; deux problèmes se posent : sélection et intrapolation. Supposons que la réaction qui a déclenché l'enregistrement a produit 4 particules émergentes qui ont traversé les plans 1 à 5 : pour reconstruire les 4 trajectoires, il faut d'abord choisir (sélection) 4 systèmes $\{P_1, P_2, P_3, P_4, P_5\}$ associant cinq impacts dans chacun des cinq plans ; et le problème se complique encore parce que certaines particules émergentes évitent un plan de détection et qu'il y a des impacts parasites. Puis le quintuplet des impacts $\{P_1, \ldots, P_5\}$ étant choisi, il faut reconstituer la trajectoire, (intrapolation), principalement pour calculer le quadrimoment de la particule.

Voici comment, selon une méthode conçue par H. Wind depuis plus de 4 ans, et réexposée par lui au colloque (Genève 1976) ; ces problèmes sont résolus par l'analyse factorielle. Une trajectoire peut être caractérisée complétement par cinq paramètres, qui sont par exemple, les coordonnées (x_1, y_1) et (x_2, y_2) de ses impacts dans les plans 1 et 2, et la quantité de mouvement mv qui détermine la courbure dans la zone du champ et donc les trois autres impacts (x_3, y_3), (x_4, y_4), (x_5, y_5). Dans l'espace R^{10} des quintuplets d'impacts $\{(x_i, y_i) \mid i = 1, \ldots, 5\} \in R^{10}$, chaque quintuplet a 10 coordonnées) les quintuplets qui peuvent correspondre à une trajectoire physiquement réalisable forment donc une sous-variété de dimension 5 : $V^5 \subset R^{10}$. Donc la sélection des quintuplets consiste à accepter seulement ceux qui sont dans V^5 ; et l'intrapolation (e.g. le calcul de mv) est un problème de régression des paramètres inconnus en fonction des coordonnées locales de V^5.

Cet énoncé géométrique n'avancerait en rien la solution si la variété V^5 n'avait une forme très simple : en fait, elle est à peu près plate. Partons d'un échantillon fictif I de 1000 trajectoires, choisies pour représenter la diversité des corpuscules susceptibles de traverser le spectromètre (divers angles, divers écarts à l'axe, divers moments mv) : les quintuplets de points qui leur correspondent forment dans R^{10} un nuage N(I) entièrement porté par V^5. L'analyse factorielle (analyse en composantes principales, i.e. recherche dans R^{10} des axes principaux d'inertie et des moments d'inertie du nuage N(I)), fournit cinq premières valeurs propres nettement séparées des cinq dernières. Si on rapporte l'espace R^{10} à 10 nouvelles coordonnées $\{t_1, \ldots, t_{10}\}$ comptées sur le

système orthonormé des axes factoriels, le quintuplet des impacts d'une trajectoire pourra être caractérisé par la propriété que la somme $\{t_6^2 + t_7^2 + t_8^2 + t_9^2 + t_{10}^2\}$ des carrés des cinq dernières coordonnées est négligeable. De façon précise on voit sur l'histogramme de la fig 4-7 que les valeurs prises par $\Sigma\{t_\alpha^2 \mid \alpha = 6, \ldots, 10\} = d^2$, pour une trajectoire et pour une combinaison erronée d'impacts (quintuplets de points mêlant les impacts de plusieurs trajectoires) se séparent bien. Ainsi la quantité critère d^2 résout le problème de la sélection ; et une régression par rapport aux 5 premiers facteurs (coordonnées t_1 à t_5) permet de calculer mv/e. Il faut cependant noter qu'il serait impossible de soumettre à la sélection par le critère d^2 tous les quintuplets d'impacts ; mais on choisit d'abord (présélection) ceux qui en projection sur un plan parallèle au champ (plan des y z sur la figure 4-6) s'alignent à peu près; puis on achève le tri par le critère d^2.

Le succès remarquable obtenu par l'analyse en composantes principales dans la sélection et l'intrapolation des trajectoires du spectromètre R603, s'explique par la linéarité de la variété V^5. En général pour un spectromètre de configuration quelconque, les coordonnées des impacts d'une trajectoire définissant un point de R^n (n = 12 s'il y a 6 plans et non 5 comme dans le R603 ; etc.) et les combinaisons acceptables qui comme les trajectoires dépendent de 5 paramètres, définissent une sous-variété $W^5 \subset R^n$; mais W^5 n'est pas plate ; et le changement

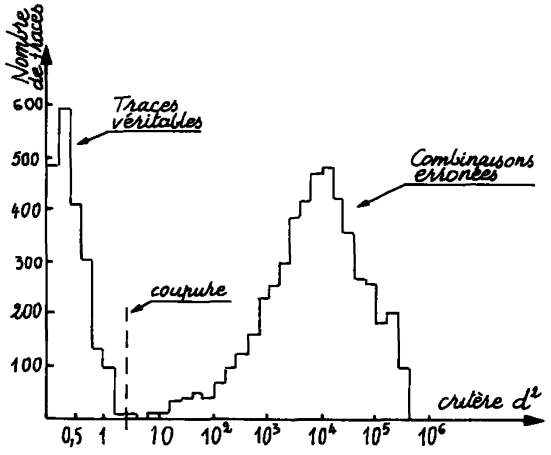

Figure 4-7 Sélection des combinaisons d'impacts correspondant à une trace véritable ; cf CERN, DD/73/23 (*) & DD/73/31 (**) ; sur l'histogramme, les valeurs élevées du critère d^2 obtenues pour des combinaisons erronées se séparent nettement.

(*) *H. Grote, M. Hansroul, J.C. Lasalle, P. Zanella : Identification of digitized particle trajectoires : CERN ; DD/73/23 ; 1973).*

(**) *M. Hansroul, D. Townsend, P. Zanelle :* The application of multidimensional analysis techniques to the proceecing of event data from large spectrometers ; *CERN ; DD/73/31 ; (1973).*

linéaire(*) de coordonnées dans R^n que fournit l'analyse factorielle permettra seulement d'avoir les 5 premiers axes bien orientés relativement à W comme sur la figure 4-8 ; on pourra tenter d'exprimer les coordonnées suivantes en fonction des cinq premières : e.g. $t_6 = f(t_1, t_2, t_3, t_4, t_5)$.

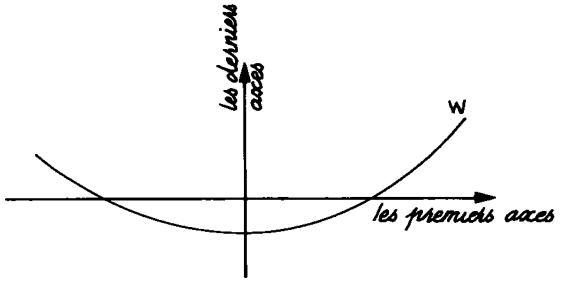

Figure 4-8 La variété W se projette convenablement sur la sous variété linéaire engendrée par les premiers axes factoriels

Cependant nous n'avons rien dit des traces incomplètes (trajectoires qui évitent un plan de fil ; ou impacts perdus accidentellement) qui correspondent à des combinaisons de points n'ayant pas en tout, les n coordonnées postulées dans le modèle $W \subset R^n$. Ici une voie - explorée présentement par J. Kubler - nous semble être de traiter un détecteur (plan de fils ou autres) non comme un instrument recueillant des nombres réels (coordonnées) mais comme une question d'un questionnaire à laquelle la réponse de la particule serait soit l'omission soit l'impact dans une des zones du détecteur (divisé e.g. en 16 carrés). Ce codage sous forme logique des informations recueillies donne un format uniforme aux trajectoires logiques qu'elles soient ou non incomplètes, et il se prête à l'analyse des correspondances.

4.4.2 *L'analyse des fluctuations dans un événement* : W. Kittel (Prague 1974 ; et exposé oral à Genève (1976) cite plusieurs travaux dont l'objet est de déceler au sein de la gerbe des particules émergentes d'un événement, une tendance à s'agréger en classes (généralement les deux classes J_1 et J_2 - avant et arrière - qui s'imposent pour un événement longitudinal : cf § 4.2.2, fig 4-3). Soulignons qu'il ne s'agit pas de répartir en classe un échantillon d'événements, mais de préciser la structure de chaque événement (ce qui toutefois n'est fait qu'en confrontant les événements entre eux.

Dans l'analyse des fluctuations de Ludlam & Slansky (cf *Phys. Rev.* D8 (1973) 1408) on considère pour chaque particule émergente une grandeur y (e.g. le moment longitudinal p_L, cf 4.2.2). Sur la totalité des particules émergentes formées dans un ensemble d'événements, on peut déterminer la loi globale $p_y = p(y)\,dy$ de y, considéré comme une variable

(*) *Des généralisations non-linéaires de l'analyse factorielle intéressent légitimement certains chercheurs (cf e.g., au programme du colloque de Genève : N. Manton :* A non- linear mapping technique in multidimensional space ; *et aussi J. Friedman :* Determination of non-linear relationships in multidimensional data). *Nous croyons toutefois que les transformations linéaires, et la recherche des axes principaux d'inertie d'un nuage, sont des opérations très efficaces auxquelles il importe de recourir même pour construire des formules non linéaires. On pourra, e.g., conjuguer analyse factorielle et régression non-linéaire, ou soumettre les données à un codage logique préalable (comme on le propose ici pour les détecteurs) ; cf § 4.5* in fine

aléatoire. Dès lors un événement unique à n particules émergentes apparaît comme un échantillon d'effectif n de cette loi p_Y. Or bien que la moyenne de ces échantillons s'identifie à la loi p_Y, il se peut que ceux-ci ne présentent pas individuellement les caractères d'un échantillon obtenu par n tirages successifs indépendants mais apparaissent comme des fluctuations s'écartant anormalement de p_Y (un cas extrême, très facile à concevoir est celui où les particules émergentes d'un événement sortiraient deux par deux avec le même quadrimoment ; un cas réel est plutôt qu'une gerbe de mésons π se divise en deux sous-gerbes bien concentrées en vitesse. L. & S. calculent un indice de fluctuation, mesurant l'écart entre un événement et la loi p_Y ; et ils comparent la distribution empirique de cet indice sur un ensemble d'événements, à ce qu'elle serait théoriquement dans le cadre d'un modèle ne comportant pas d'agrégation systématique (sous-gerbes) au sein des événements individuels.

Par de tels calculs, on peut conclure au rejet de l'hypothèse nulle, c'est-à-dire à l'existence d'une certaine agrégation systématique, mais non préciser en quoi consiste celle-ci. Pour reconnaître un agrégat (sous-gerbe) au sein d'un événement, Berger, Fox & Kryzwicki (cf *Phys. letters* 43 B (1973) 132, proposent d'écarter la particule émergente la plus excentrique, puis de calculer si la variance est anormalement basse ; (anormalement, ici encore, par rapport à un modèle de référence sans agrégation). Le procédé de B. F. & K. nous paraît insuffisamment général, en ce qu'il n'envisage que l'existence d'un agrégat formé de n - 1 particules émergentes, une seule étant écartée ; de plus les calculs, comme ceux de L. & S. portent sur une variable unidimensionnelle y. Il est possible d'étudier en toute généralité l'ensemble des sous-agrégats qu'on peut former au sein d'un événement et d'en critiquer la validité. Voici comment.

Pour fixer les notations, disons qu'il y a 10 particules émergentes e_j à chacune desquelles est associée une grandeur vectorielle y_j prenant ses valeurs dans un espace euclidien E. Soumettons $\{y_j \; j = 1, \ldots, 10\}$ à la classification ascendante hiérarchique (cf § 4.3.4) avec agrégation suivant la variance. (i.e. en choisissant de réunir deux classes c et c' déjà formées selon le critère que la variance $c \cup c'$ dépasse le moins possible la somme des variances de c et c') : l'arbre de classification obtenu a pour éléments terminaux les dix y_j, et il comprend 9 noeuds, chacun représentant une classe ayant un certain effectif p_c, et une certaine variance v_c. Dans le plan, l'ensemble des points (p_c, v_c) présente une suite de raies d'abscisse 2,...,10, parallèles à l'axe des v : le nuage obtenu par dépouillement de données réelles peut être comparé à celui d'un modèle de référence : et l'on pourra décider d'accepter comme agrégats valides les classes c pour lesquelles, à la valeur p_c de l'effectif, la variance v_c est anormalement basse.

Dans une recherche en cours (G. Fontaine et D. Maïti) on considère des événements ayant produit une particule à fort moment transverse ; ce qui suggère clairement une structure en quatre jets : deux jets longitudinaux (héritiers de deux protons dont le choc a produit l'événement) ; un jet transversal dominé par la particule à fort moment transverse ; et encore un jet approximativement opposé à celui-ci pour que soit respectée la conservation du quadrimoment. Ces recherches sont d'autant plus difficiles que les particules non-chargées ne sont pas détectées (données dites *inclusives*).

4.4.3 <u>*Validité de l'ajustement*</u> : L'analyse des fluctuations dans un événement nous a conduit à concevoir par simulation une épreuve de validité pour les agrégats de particules émergentes. De telles épreuves sont les seules possibles dans la pratique de l'analyse des données : car, les modèles simples ne se rencontrent guère, les formules élémentaires et les lois tabulées sont sans objet. En physique corpusculaire, la solidité des lois (cf § 3.0) impose à l'analyse des données un cadre

géométrique (l'espace des phases, cf §§ 3.1 & 3.2) qui n'a pas d'équivalent dans les sciences humaines. Et des formules analytiques semi-empiriques servent pour découvrir la structure des canaux (§ 4.1). Discuter de la validité de ces formules est *a priori* inutile, puisque même si elles s'ajustent aux données, nul ne songe à affirmer qu'elles soient des lois physiques valides. Mais, sans s'attacher à des hypothèses semblables à celles de la statistique paramétrique, on peut souhaiter éprouver la validité de l'ajustement d'une formule à un échantillon d'événements. Cependant, en statistique multidimensionnelle, la comparaison directe d'une formule de loi à un échantillon (nuage multidimensionnel) est généralement impraticable ; ce qui des formules analytiques nous ramène, ici encore, aux épreuves de simulation. Pour comparer à une loi p_E un nuage I_1 inclus dans un espace E, on engendre par simulation un nuage I_2 issu de la loi p_E et on cherche à éprouver l'hypothèse qu'aux deux nuages I_1 et I_2 correspond dans l'espace E une même densité. Ce dernier problème intéresse lui-même la physique expérimentale : car on pourra e.g. se demander si deux échantillons I_1 et I_2 produits dans des conditions différentes diffèrent significativement. Au colloque de Genève les méthodes non paramétriques d'ajustement multidimensionnel furent présentées par B. Schorr (et aussi M.J. Grimon) ; ne disposant pas de notes sur ces exposés, nous nous reporterons à un travail antérieur de J. Friedman, publié au CERN(*).

Pour comparer les lois de I_1 et I_2 , on unit ces deux échantillons d'effectifs respectifs N_1 et N_2 , en un échantillon unique I d'effectif N. Pour chaque point i de I, on considère le contenu de la boule B(i,R) de rayon R, de centre i, contenant exactement k individus de I (e.g. k=20; cf § 4.3.2). Parmi ces k individus, il y en a $k_1(i)$ de la classe I_1 ; et $k_2(i) = k - k_1(i)$, de la classe I_2. En bref $(k_1(i)/N_1)/(k_2(i)/N_2)$ est un estimateur du rapport en i des densités des lois dont sont issus les deux nuages. Si ces deux lois diffèrent grandement, $k_1(i)$ (dont la valeur moyenne est kN_1/N) variera avec i plus qu'on ne peut l'attendre du jeu des fluctuations d'échantillonnage ; des valeurs soit très faibles ($k_1(i)=0$, $k_1(i) = 1$), soit très élevées ($k_1(i) = k$, $k_1(i) = k - 1$) seront atteintes selon que i appartient au domaine où prédomine I_2 , ou à celui où prédomine I_1. De façon précise, on construit d'une part l'histogramme H des valeurs réparties en k + 1 classes de 0 à k) correspondant à l'échantillon I_1 donné ; et d'autre part plusieurs histogrammes analogues H', H" , H"' ,..., obtenus sur des échantillons fictifs I'_1 , $I"_1$, $I"'_1$ construits en tirant au sort (Monte-Carlo) N_1 des points de I : et l'on éprouve si l'histogramme H rentre dans la distribution de dispersion des histogrammes simulés H', H", H"' ,autour de leur histogramme moyenne (qui est en première approximation, une loi binomiale).

Par cette méthode J. Friedman, (in *Phys. Rev.* D, Vol 9, pp 3053-3059; (1974)), a notamment trouvé que pour la réaction $p\,p \to p\,p\,\pi^+\pi^+\pi^-\pi^-$, il n'y a pas de différence significative dans la distribution des moments transverses, entre deux échantillons de quelque 200 événements chacun , pour lesquels l'énergie du proton incident vaut respectivement 12 GeV et 28 GeV.

Avec les épreuves de validité multidimensionnelle, J. Friedman (CERN, Godøysund, 1974) expose la recherche des corrélations. Il pose le

(*) Data analysis techniques for high energy particle physics ; in *Proceedings of the 1974 CERN school of computing* ; Godøysund, Norvège; CERN/74/23 ; 1974).

problème suivant : étant donné un échantillon de points (x,y) appartenant à un espace produit X x Y, dire si la loi p_{XY} de l'échantillon est une loi produit $p_X\ p_Y$; et il passe du problème continu à un problème fini en partageant X et Y en des ensembles de cellules que nous noterons respectivement I et J :

$$X = \cup \{X_i \mid i \in I\} \ ; \quad Y = \cup \{Y_j \mid j \in J\}.$$

Si l'on note k(i,j) le nombre des points (x,y) appartenant au produit $X_i\ Y_j$ de la cellule i par la cellule j, on aboutit au problème usuel de l'analyse des correspondances : J.F. ne traite pas à fond ce problème de la recherche des corrélations, il se borne à discuter par un critère issu de la théorie de l'information, la validité de l'hypothèse nulle d'indépendance entre certaines variables prises deux à deux ; ce qui prépare la recherche ultérieure des corrélations par diverses méthodes analytiques (cf e.g. § 4.1).

4.4.4. Conclusion : rôle de la statistique : Histogrammes et diagrammes triangulaires ont permis de découvrir les résonances corpusculaires (ou états de particule de durée de vie infiniment petite ; cf N*, K*, § 3.5). Mais selon W. Kittel (Prague 1974) des sous-canaux ou des interférences indistincts sur les représentations de dimension 1 ou 2 se détachent à l'évidence par l'analyse multidimensionnelle d'échantillons d'effectif modeste ; et W. K. conclut non sans humour :

"On se demandait s'il valait la peine de traiter statistiquement les processus pluricorpusculaires *(Should one trent multiparticle processus?)* Il est aujourd'hui acquis que la réponse doit être affirmative. Mais une autre question se pose alors : comment traiter *(How could one treat multiparticle processes)*. A cette dernière question, il n'est pas si facile de répondre qu'à la première ; mais quoiqu'il en soit des méthodes particulières requises, nous dirons en tout cas : traitez-les avec le plus grand soin ! *(it sould be with greatest possible care)*".

On nous permettra de ne pas nous satisfaire de ce conseil. Il ne fait pas de doute qu'outre de la prudence dans l'emploi des méthodes statistiques, l'audace dans la recherche des lois analytiques de la physique est requise pour voir et comprendre la forme des nuages d'événements. Car il ne s'agit pas ici de vagues classes, branches ou ilôts *(clusters)*, mais au fond, des surfaces de niveau du module de certaines fonctions analytiques qui, on ne sait pas au juste encore comment, relient les mutiples formes de la matière.

4.5 Appendice : L'estimation robuste en analyse multidimensionnelle:

Reprenons dans le langage géométrique qui nous est familier, l'exposé que P. Huber a donné de ses travaux et de ceux de plusieurs autres auteurs.

On sait que dans le calcul du centre de gravité G et de la matrice σ des variances-covariances d'un nuage de points, quelques individus excentriques (en anglais : *outlyers*) vraisemblablement étrangers à la population principale étudiée, peuvent avoir une influence considérable dont on désire s'affranchir. Pour cela on part d'une définition de G et σ par le maximum de vraisemblance, associée au modèle normal multidimensionnel ; puis on généralise cette définition en substituant à la loi normale un autre modèle possédant la symétrie sphérique.

Plaçons-nous dans R^p, rapporté à un système de coordonnées $x_1, \ldots, x_j, \ldots, x_p$. Si R^p est muni de la norme $|x| = (\Sigma x_j^2)^{1/2}$, la loi normale sphérique ayant variance 1 dans toutes les directions s'écrit :

$$(2\pi)^{-p/2} \exp(-|x|^2/2)\, dx_1 \ldots dx_p \ .$$

Une loi normale spatiale quelconque sur R^p, peut être complétement définie par la donnée de son centre G et d'une métrique euclidienne $\|x\|_n$

définie par une matrice $n = \{n^{jj'} | j,j'=1,\dots,p\}$ qui n'est pas en général la matrice diagonale unité ; avec la métrique :

$$\|x\|_n^2 = \Sigma\{n^{jj'} x_j x_{j'} \mid j = 1,\dots, p \ ; \ j' = 1,\dots, p\},$$

on a donc la loi normale $N_{G;n}(x)$:

$$(2\pi)^{-p/2} \exp(-\|Gx\|_n^2/2) \det(n)^{1/2} dx_1 \dots dx_n = N_{G;n}(x)\, dx \ ;$$

et la matrice des variances-covariances de cette nouvelle loi normale n'est autre que la matrice σ inverse de la matrice n : $\sigma = n^{-1}$.

$$\int_{R^p} (x_j - G_j)(x_{j'} - G_{j'}) \, N_{G;n}(x)\, dx = \sigma_{jj'} .$$

Soit maintenant R^p un nuage $\{M_i \mid i \in I\}$, où l'on suppose que les points $M_i = \{M_{i1}, \dots, M_{ij}, \dots, M_{ip}\}$ ont tous même masse. Du point de vue du maximum de vraisemblance, on associe à ce nuage la loi normale $N_{G;n}$ rendant maximum le produit :

$$\Pi\{N_{G;n}(M_i) \mid i \in I\} \ ;$$

il se trouve que ce maximum est atteint si l'on prend pour G le centre de gravité du nuage et pour n l'inverse de sa matrice des variances-covariances : $n = \sigma^{-1}$;

$$G_j = (1/\mathrm{Card}\, I)\, \Sigma\{M_{ij} \mid i \in I\}$$

$$\sigma_{jj'} = (1/\mathrm{Card}\, I)\, \Sigma\{(M_{ij} - G_j)(M_{ij'} - G_{j'}) \mid i \in I\}.$$

Cette propriété peut être retournée en une définition du centre de gravité et de la matrice des variances-covariances. On appellera centre de gravité d'un nuage le centre de la loi normale ajustée à celui-ci par le maximum de vraisemblance ; et matrices des variances-covariances du nuage, la matrice n^{-1} de cette loi.

Soit maintenant une fonction positive $f(\rho)$ définie pour $\rho \in (0,\infty)$ et telle que :

$$\int_{R^p} f(|x|)\, dx_1 \dots dx_p = 1 \ ;$$

la fonction f fournit un modèle de loi de probabilité possédant la symétrie sphérique ; le cas de la loi normale correspondant à $f(\rho) = (2\pi)^{-p/2} \exp(-\rho^2/2)$. Par transformation affine, on obtient à partir de la loi associée à f, une famille de lois $\mathcal{F}_{G;n}$, dont chacune est donnée par son centre G et une métrique euclidienne $\|x\|_n$ suivant la formule :

$$f(\|Gx\|_n) \det(n)^{1/2}\, dx_1 \dots dx_p = \mathcal{F}_{G;n}(x)\, dx.$$

A un nuage $\{M_i \mid i \in I\}$, peut être associé suivant le principe du maximum de vraisemblance la loi $\mathcal{F}_{G;n}$ rendant maximum le produit :

$$\Pi\{\mathcal{F}_{G;n}(M_i) \mid i \in I\} \ ;$$

Le centre G de cette loi pourra être appelé centre généralisé (ou : f-centre) du nuage ; et $\sigma = n^{-1}$, sera la matrice des variances-covariances généralisées (ou f-matrice). Du point de la stabilité, ou robustesse, l'intérêt de cette définition est qu'il est possible de la modifier en sorte que soit réduite l'influence des éléments les plus excentriques du nuage : ici se poseraient aux mathématiciens maints problèmes difficiles d'existence et d'unicité.

A cette procédure élégante, et qui vise certes un problème dont l'importance est réelle, nous ferons plusieurs objections ; et conclurons en présentant la solution pragmatique que nous avons adoptée en analyse des données.

Dans son principe, la procédure est fondée sur l'invariance affine, qui respecte la classe des lois N ou $\mathcal{F}$. Cependant dans la plupart des problèmes concrets, l'espace ambiant au nuage a une structure métrique, et c'est sur la considération simultanée de deux formes quadratiques: la forme quadratique d'inertie σ du nuage et la métrique m de l'espace ambiant (qui n'est qu'un $n = \sigma^{-1}$), que repose l'étude d'un nuage (cf TII B n° 2).

De même qu'il y a des points parasites M_i, il y a des variables parasites x_j (ceci est bien clair en analyse des correspondances, où individus et variables jouent des rôles symétriques). Ecarter les variables parasites de l'espace ambiant, se fait généralement en projetant le nuage sur les sous-espaces engendrés par les premiers axes factoriels ; axes dont la détermination utilise essentiellement la métrique m de l'espace ambiant.

En s'efforçant, moyennant le choix d'une métrique convenable n, (ou ce qui est équivalent, par une transformation affine) d'ajuster au nuage un modèle possédant la symétrie sphérique, on fait grandement violence aux données si le nuage est bimodal, voire multimodal.

Aussi sans méconnaître les formulations mathématiques générales, nous nous en tenons à la pratique suivante : considérer le nuage $\{M_i | i \in I\}$ dans les plans engendrés par les premiers axes factoriels (axes 1 x 2, 2 x 3 etc) et mettre en éléments supplémentaires (de masse évanescente ; donc se plaçant sur les axes mais ne jouant aucun rôle dans la construction de ceux-ci) les points excentriques, et plus généralement tout point qui (soit par son excentricité, soit par sa forte masse) apporte à un axe α une contribution représentant une part élevée (e.g. 20%) du moment d'inertie correspondant λ_α (en bref : $\lambda_\alpha = \Sigma$ masse(i) $F_\alpha(i)^2$ et masse(i) x $F_\alpha(i)^2$ est la contribution de i à λ_α). Par de telles précautions, on parvient à des résultats robustes, pour autant que les données recueillies le permettent. Comme partout en analyse des données, le modèle linéaire et euclidien (moyennes, variances, matrices d'inertie) s'impose ici par la grande simplicité des calculs : plutôt que de s'aventurer dans un cadre plus général mais requérant une puissance de calcul que nous n'avons pas, il est généralement préférable de se borner à retoucher (*) les données traitées et les résultats de calcul.

(*) *En particulier, le codage sous forme logique des informations numériques continues (cf § 4.4.1* in fine), *tout en produisant un tableau ordinaire de nombres qu'on analyse suivant le modèle linéaire de l'analyse factorielle, introduit des fonctions en créneaux des variables initiales. C'est là une généralisation qui dépasse les calculs des polygones algébriques. De plus, donner aux valeurs extrêmes des données une représentation (binaire ; en 0,1) qui ne diffère pas de celle des valeurs ordinaires les plus grandes - ni des plus petites - (car dans les modalités d'une variable telles que très fort et très faible, les individus ordinaires sont la majorité ; les cas extrêmes sont l'exception), on réduit radicalement le rôle des individus aberrants. Pour un exemple de ce traitement cf* [Erodium] *Cahier Vol II n° 1, pp 97-113.*