LES CAHIERS DE L'ANALYSE DES DONNÉES

P. LUTZ
D. MAÏTI

Classification automatique d'après la distance entre orbites : application à la physique corpusculaire

Les cahiers de l'analyse des données, tome 3, n° 4 (1978), p. 449-458

<http://www.numdam.org/item?id=CAD_1978__3_4_449_0>

Les Cahiers de l'Analyse des Données
Vol. III - 1978 - n° 4 - p. 449-458

CLASSIFICATION AUTOMATIQUE D'APRÈS LA DISTANCE ENTRE ORBITES : APPLICATION A LA PHYSIQUE CORPUSCULAIRE [DIST. ORB. PHYS. COR.]

par P. Lutz et D. Maïti (1)

La présente note, destinée à des lecteurs non spécialistes de physique, donne d'abord les principes géométriques (§ 1) des algorithmes de classification (§ 2) ; puis esquisse un cas de données physiques (§ 3). Tout au long de la note, intervient la recherche du déplacement minimisant la distance entre deux ensembles de points homologues. Ce dernier *problème* (cf *supra* dans ce cahier) a été en fait suggéré par des études de physique : d'abord celles de G. Flamembaum sur les conformations de molécules , puis celles présentées ici relatives aux réactions de physique corpusculaire à haute énergie ; mais dans l'exposé qui suit, nous supposons que le lecteur a présent à l'esprit, l'essentiel du *problème*.

1 *Distance entre orbite* : Nous présentons sommairement ici des notions bien connues de géométrie.

1.1 *Définition des orbites* : En mécanique, on appelle orbite l'ensemble des positions successivement occupées par un point mobile ; ou encore l'ensemble des positions qu'est susceptible d'occuper un point d'un corps solide assujetti dans son déplacement, à certaines liaisons etc . Ici on donne à orbite le sens suivant. Soit F un espace métrique (i.e . espace entre les points duquel est définie une distance Dis satisfaisant à l'inégalité du triangle) ; G un groupe d'isométries de F ; i.e. un sous-groupe du groupe de toutes les applications biunivoques de E dans E, satisfaisant à la condition de respecter la distance :

$$\forall\, g \in G\ ,\ \forall\, M, M' \in F\ :\ Dis(gM, gM') = Did(M, M')\ ;$$

on appelle orbite d'un point M de F par les transformations du groupe G, l'ensemble de tous les transformés de M par une isométrie g de G ; ce que l'on note :

$$Orb(M) = GM = \{ gM \mid g \in G \}$$

On remarque que tout point M' = gM, transformé de M par une isométrie g G, a même orbite que M : en fait les orbites ne sont autres que les classes d'équivalences de la relation

$$M \approx M' \Leftrightarrow \exists\, g \in G : gM = M'$$

or si gM = M', on a aussi $g^{-1}M' = M$ et la transformation g^{-1} inverse de g est dans G, puisque G est un groupe. Dans la figure, on a représenté les orbites comme des courbes ou filtres ; l'ensemble de ces orbites est

(1) *Laboratoire de physique corpusculaire, Collège de France. Thèse de P. Lutz : Contribution à la mesure automatique des clichés de chambre à bulles et à l'analyse multidimensionnelle des interactions K^-p à 14,3 GeV.*

Thèse de D. Maïti : Contribution de l'analyse factorielle et de la classification automatique à l'étude de la réaction $K^-p \to K^-p\,\pi^+\pi^-$ à 14,3 GeV.

appelé l'ensemble quotient F/G de l'espace F par les opérations du groupe G. Si l'on fait abstraction des singularités, on imaginera que F est

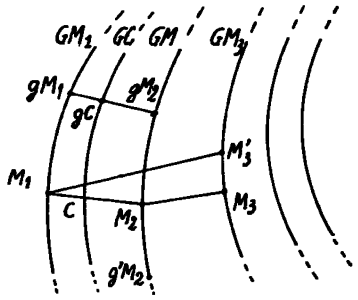


Figure : Orbites et distances entre orbites : on considère particulièrement les trois orbites GM_1, GM_2, GM_3 : pour illustrer la démonstration de l'inégalité du triangle (cf § 1.3) ; et l'orbite moyenne GC entre GM_1 et GM_2 (cf § 2.2).

un espace de dimension d , G un groupe dont les performances dépendent des paramètres ; une orbite ou filtre sera en général une variété de dimension r (mais il peut y avoir des points fixes !) ; et l'ensemble quotient est de dimension d - r : on peut le représenter (au moins localement) comme une sous-variété de F transverse aux orbites.

1.2 *Exemple de nuages de points homologues* : Le problème et l'article qui précèdent, fournissent l'exemple d'orbites que nous avons en vue . Notons donc comme dans le problème E un espace euclidien (dim E = n) ; I un ensemble fini (Card I = m) ; $\{m_i \mid i \in I\}$ un système fini de nombres positifs (masses) indicé par $i \in I$. L'ensemble des systèmes $M(I) = \{M^i \mid i \in I\}$ de points de E, indicés par $i \in I$, peut être considéré comme un espace $F = E^m$ de dimension $d = m \times n$ (chaque système M(I) est caractérisé par les m n coordonnées de ses m points : n coord. par point). Entre deux systèmes M(I) et N(I) est définie une distance Dis suivant la formule (du problème) où interviennent les masses m_i :

$$\mathrm{Dis}(M(I), N(I)) = \Sigma\{m_i \|M^i - N^i\|^2 \mid i \in I\} ;$$

Cette formule fait de F un espace métrique, plus précisément un espace euclidien (on voit en effet que Dis est une somme pondérée par les m_i des différences aux carrés des coordonnées. Les points M^i , N^i ; i.e. des coordonnées de M(I) et N(I) considérés comme éléments de F).

Soit maintenant G le groupe des isométries de l'espace euclidien E (espace ambiant aux systèmes de points M(I) , N(I)) ; (éventuellement on prendra seulement pour G un sous-groupe du groupe des isométries de E : e.g. les isométries directes ; les rotations autour d'un axe ; etc ; cf *infra* § 3) ; un élément $g \in G$ définit aussi une isométrie (notée également g) de l'espace F des systèmes de points ; il suffit de poser :

$$gM(I) = \{gM^i \mid i \in I\} ;$$

on a bien une isométrie car :

$$\forall\, g \in G : \mathrm{Dis}(M(I), N(I)) = \mathrm{Dis}(gM(I), gN(I)) ;$$

$$\Sigma\{m_i \|M^i - N^i\|^2 \mid i \in I\} = \Sigma\{m_i \|gM^i - gN^i\|^2 \; i \in I\}.$$

1.3 *Définition de la distance entre orbites* : Dans le *problème*, on cherche à déplacer le nuage N(I) en sorte qu'il s'écarte le moins possible du nuage M(I). Dans le langage de la présente note nous dirons qu'on cherche sur l'orbite GN(I) (ensemble des nuages qu'on peut obtenir en déplaçant N(I) par $g \in G$) un élément gN(I) (dans l'article de Ph. Bourgeois, il est noté αN(I)) réalisant le minimum de la distance à M(I). En général ce minimum sera appelé la *distance orbitale* Dor entre deux points M et N de E ; celle-ci est définie par la formule :

$$\mathrm{Dor}(M,N) = \inf\{\mathrm{Dis}(M,gN) \mid g \in G\} ;$$

Il est clair que la distance Dor(M,N) dépend seulement des orbites GM et GN, non des éléments particuliers M et N considérés sur ces orbites : Dor(M,N) est le minimum de la distance Dis entre GM et GN, i.e. entre un point $M' = g'M \in GM$, et un point $N'' = g''N \in GN$. Ainsi Dor est défini sur l'espace quotient F/G. Est-ce une distance? Rappelons les axiomes d'une distance :

symétrie : $\forall\, M,N$: Dor(M,N) = Dor(N,M) : cette propriété est évidente sur la définition

positivité : $\forall\, M,N$: $\mathrm{Dor}(M,N) \geq 0$;

$$\mathrm{Dor}(M,N) = 0 \Leftrightarrow GM = GN.$$

il est clair que Dor est positif ou nul comme Dis ; en revanche la condition : "si la distance Dor entre deux orbites est nulle celles-ci coïncident" n'est pas nécessairement satisfaite ; il se peut que deux orbites approchent arbitrairement près l'une de l'autre (inf = 0) sans pour autant avoir de point en commun (ce qui entrainerait leur identité). Toutefois dans le cas que nous avons en vue (cf §3) il est facile de voir que si N et M ne sont pas sur une même orbite, $\mathrm{Dor}(M,N) \neq 0$.

inégalité du triangle :

$$\forall\, M_1, M_2, M_3 : \mathrm{Dor}(M_1, M_3) \leq \mathrm{Dor}(M_1, M_2) + \mathrm{Dor}(M_2, M_3).$$

On démontrera cette inégalité d'après la figure. Pour simplifier, supposons que le minimum de la distance entre M_1 et GM_2 est précisément réalisé entre M_1 et M_2 ; et de même que le minimum de la distance entre M_2 et M_3 est réalisé entre M_2 et M_3 :

$$\mathrm{Dor}(M_1, M_2) = \mathrm{Dis}(M_1, M_2) \ ; \ \mathrm{Dor}(M_2, M_3) = \mathrm{Dis}(M_2, M_3) \ ;$$

puisque la distance Dis (donnée sur F) satisfait à l'inégalité du triangle on a :

$$\mathrm{Dis}(M_1, M_3) \leq \mathrm{Dis}(M_1, M_2) + \mathrm{Dis}(M_2, M_3) = \mathrm{Dor}(M_1, M_2) + \mathrm{Dor}(M_2, M_3) \ ;$$

or la distance $\mathrm{Dor}(M_1, M_3)$, minimum de la distance entre M_1 et un point de GM_3 (minimum réalisé sur la figure par $\mathrm{Dis}(M_1, M'_3)$ est nécessairement inférieur ou égal à $\mathrm{Dis}(M_1, M_3)$, d'où l'inégalité du triangle.

$$\mathrm{Dor}(M_1, M_3) \leq \mathrm{Dis}(M_1, M_3) \leq \mathrm{Dor}(M_1, M_2) + \mathrm{Dor}(M_2, M_3).$$

2 *Agrégation suivant la distance orbitale* : On se propose de soumettre à la classification automatique, un ensemble fini $\mathrm{Ens} \subset F$, en se fondant non sur la distance primaire Dis, mais sur la distance Dor : ce qui revient encore à appliquer la classification à l'ensemble d'orbites $G\,\mathrm{Ens} = \{GM \mid M \in \mathrm{Ens}\} \subset F/G$. Pratiquement toutefois les calculs se font non dans le cadre de l'espace quotient F/G, mais dans F lui-même, en modifiant s'il est nécessaire les algorithmes usuels.

2.1 Diversité des algorithmes d'agrégation : Nous considérerons seulement deux algorithmes : la classification ascendante hiérarchique (CAH), et l'agrégation autour des centres variables (E. Diday : Nuées dynamiques).

On sait qu'en CAH, on agrège d'abord les deux éléments M,M' de Ens les plus proches ; et l'on poursuit en agrégeant entre eux deux à deux soit les éléments restants soit les classes déjà constituées ; l'algorithme pouvant être appliqué avec de nombreux critères d'agrégation entre classes. Certains de ces critères : agrégation suivant le saut minimum, agrégation suivant le diamètre, agrégation suivant la distance moyenne, font seulement usage des distances données initialement entre éléments M,N,.. de Ens : il importe peu que ces distances aient été calculées suivant la formule de Dis(M,N) ou celle de Dor(M,N); les propriétés de la formule n'interviennent pas dans les calculs ultérieurs.

Mais la procédure d'agrégation suivant la variance est conçue dans le cadre d'un espace euclidien : en bref le critère d'agrégation entre deux classes a et b (on peut dire pour faire image : le *coût* de l'imprécision commise en confondant ces deux classes) est donnée par la formule :

$$Dcrit(a,b) = (m_a\ m_b/(m_a + m_b))\ \|a\ b\|^2$$

(où m_a , m_b sont les masses des classes a et b, sommes des masses des individus de ces classes ; et $\|a\ b\|^2$ est le carré de la distance euclidienne entre les centres de gravité des classes) ; et Dcrit(a,b) s'interprète comme l'inertie d'un système de deux points ayant pour masses respectives m_a et m_b et séparés par la distance $\|a\ b\|$ des centres de gravités. Certes, l'algorithme général de CAH (cf TII B n° 4 ; spécialement le § 2.5.3) ne calcule pas les coordonnées des centres de gravité des classes ; il ne considère même pas les coordonnées des individus à agréger; mais calcule Dcrit(a,b) de proche en proche, à partir des distances entre individus (éléments de Ens) qui sont les seules données géométriques du programme. Toutefois appliquer le programme de CAH avec agrégation suivant la variance ne se conçoit que si la notion de centre de gravité conserve un sens dans le cadre géométrique où sont les individus de l'ensemble Ens. Ici, agréger suivant la variance avec la distance Dor, requiert qu'on ait défini au moyen l'orbite moyenne entre deux orbites (§ 2.2).

Dans l'algorithme d'agrégation autour des centres variables,on part d'un ensemble fini de centres $c_1^\circ, \ldots, c_p^\circ$ situés dans l'espace ambiant F ; et on affecte chaque individu de Ens à celui de ses centres dont il est le plus proche ; puis on recalcule de nouveaux centres $c_1^1, \ldots, c_p^1$ comme centres de gravités des classes constituées autour des centres initiaux ; on affecte alors les individus aux centres c^1 et ainsi de suite jusqu'à convergence : ici encore intervient la notion de centre de gravité.

2.2 Orbite moyenne entre deux orbites : Supposons (cf § 1.2) que F est un espace euclidien, et G un groupe d'isométries de F. Soit M_1 , M_2 deux points de F affectés respectivement des masses m_1 , m_2. Il est naturel de définir un point moyen C par la formule usuelle :

$$C = (m_1\ M_1 + m_2 g M_2)/(m_1 + m_2)$$

où l'isométrie g a été choisie dans le groupe G pour rendre minima la distance entre M_1 et gM_2 (i.e. $Dis(M_1 , gM_2) = Dor(M_1 , M_2)$: la distance orbitale est réalisée entre M_1 et gM_2). Et l'on dira que GC est l'orbite

moyenne entre les orbites GM_1 et GM_2 (affectées des masses m_1 et m_2). Cette définition est satisfaisante à condition que pour M_1 et M_2 donnés, le point gM_2 réalisant le minimum de la distance entre M_1 et l'orbite GM_2 soit unique : alors on voit aisément que l'orbite GC est bien définie et ne dépend que des seules orbites GM_1 et GM_2 (et non des points particuliers M_1 et M_2 pris sur celles-ci ; cf figure).

Il ne semble pas possible de définir univoquement un centre de gravité pour plus de deux points : car (cf § 1.3, inégalité du triangle; et figure) il n'est pas en général possible étant données trois orbites de trouver sur celles-ci des points M_1 , M_2 , M_3 tels que les trois distances Dis entre ces points réalisent simultanément les distances minima entre orbites (distances Dor). Mais si le système des points dont on cherche le centre de gravité est donné avec un ordre (numérotage) on pourra suivre cet ordre pour calculer de proche en proche un centre : en prenant d'abord la moyenne M_{12} entre les points M_1 et M_2 , puis la moyenne M_{123} entre les points M_{12} et M_3 ; etc.

2.3 Classification ascendante utilisant l'orbite moyenne : L'algorithme de CAH, avec agrégation suivant la variance, sera modifié comme suit.

1) Chaque fois qu'est créée une classe q on lui associe un centre (point de F noté a) dont les coordonnées sont explicitement calculées.

2) Le critère d'agrégation entre classes est le critère, avec la distance Dor entre les centres :

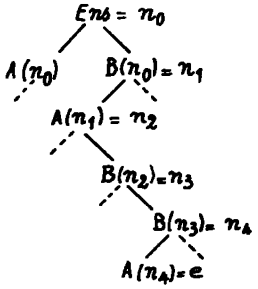
$$Dcrit(a,b) = (m_a\ m_b/(m_a + m_b))\ (Dor(a,b))^2$$

3) Quand on agrége deux classes a et b réalisant le minimum de Dcrit (à l'étape donnée du programme) le centre de la nouvelle classe $a \cup b$ est calculé comme moyenne des centres a et b comme on l'a expliqué au § 2.2. De façon précise si on note g(b;a) l'isométrie (supposée unique) tel que la distance Dor(a,b) soit réalisée comme distance Dis entre a et g(b;a)b on a :

$$a \cup b = (m_a\ a + m_b\ g(b;a)b)/(m_a + m_b)$$

(dans cette formule, $a \cup b$, a, b désignant les *centres* des classes ; m_a m_b leurs masses) ; en particulier si a et b sont des systèmes de points M(I) et N(I) (cf § 1.2), g(b;a) n'est autre que l'isométrie α déterminée explicitement par Ph. Bourgeois (isométrie qui est unique si le rang de S(M,N) est égal à la dimension de E : cf ; Bourgeois § 7.c)

Remarque : La connaissance des isométries $gn = g(B(n);A(n))$ associées à chaque noeud n de la classification hiérarchique (où on a noté A(n) et B(n) les deux descendants dits aîné et benjamin, par réunion desquels est formé le noeud $n = A(n) \cup B(n)$) permet de réaliser comme suit un applatissement de l'ensemble Ens des éléments à classer. Par applatissement, nous entendons que chaque élément $e \in Ens$ aura été soumis à une isométrie $g(e) \in G$ de telle sorte que si l'on procède à la classification sur l'ensemble des g(e)e, il ne se rencontrera jamais en aucun noeud de rotation gn autre que l'identité. Soit donc à un événement e placé comme indiqué ci-dessous la classification arborescente construite :



on posera $g(e) = gn_0 \times gn_2 \times gn_3$.

D'une façon générale si la suite des prédécesseurs de e est :

$n_0, \ldots, n_h, \ldots, n_r, n_{r+1} = e$;

on aura :

$g(e) = gn_0^0 \times \ldots \times gn_h^h \times \ldots \times gn_r^r$;

où h = 1 si $n_{h+1} = B(n_h)$ et zéro si $n_{h+1} = A(n_h)$;

i.e. on tourne seulement si n_{h+1} est le benjamin de n_h ; sinon on a un terme gn_h^0 = identité.

2.4 Agrégation des variables autour de centres variables : L'algorithme de E. Diday tel qu'on l'a esquissé au § 2.1, requiert non seulement comme CAH des calculs de moyenne entre deux points (§2.3) mais des calculs de centre de classes d'effectif quelconque. On peut (cf § 2.2) ramener les calculs de centres à des calculs de moyenne effectués de proche en proche : on peut encore modifier l'algorithme comme suit: les centres sont modifiés non seulement à chaque itération après avoir affecté tous les individus à classer, mais continuent au fur et à mesure de chaque affectation. Partant d'une suite de centres $C_1, \ldots, C_p$, on lit la suite des individus e à classer dans l'ordre où ceux-ci sont gardés en mémoire ; chaque individu e est affecté au centre dont il est le plus proche, ce centre c étant déplacé vers une nouvelle position qui est la moyenne entre l'ancienne position (muni de la masse totale de tous les individus qui lui ont déjà été attachés) et le nouvel individu e (affecté de sa masse propre).

3 *Application à l'étude des réactions entre corpuscules à haute énergie*

Après avoir rappelé le cadre physique et les notations qui lui sont propres (§ 3.0), on construit l'espace métrique F (§ 3.1 ; cf § 1.1); et propose divers choix d'un groupe G d'isométrie (§ 3.2) ; ce qui permet l'application d'algorithmes de classification (§ 3.3 : cf § 2).

3.0 Le cadre physique : L'objet élémentaire de l'étude est l'*événement*. En bref un *événement* à haute énergie est schématisé comme suit : l'interaction (on peut dire le *choc* pour faire image) de deux corpuscules incidents a et b, produit un ensemble I de corpuscules émergents . Par exemple (cf P. Lutz Thèse, & D. Maïti Thèse) dans la réaction notée :

$$K^- p \to K^- p \quad \pi^+ \quad \pi^-$$

les deux corpuscules incidents sont un méson K^- négatif, et un proton p; et la gerbe émergente comprend quatre corpuscules. Le statisticien trouvera dans les *Cahiers* une introduction à l'étude de ces réactions (cf [Phys.Cor.] : Vol II n° 3, pp 313-332 ; Vol n) 4, pp 451-466 ; Vol I n°1 pp 79-94). On rappellera seulement ici qu'à chaque corpuscule est associée une trajectoire rectiligne dans l'espace quadridimensionnel E (espace rapporté aux quatre coordonnées ct, x, y, z : où c est la vitesse de la lumière) ; et un vecteur, porté par cette trajectoire, le quadrimoment (dont les quatre composantes sont En/c , p_x , p_y , p_z ; où En = énergie). Dans les réactions entre corpuscules, il y a conservation du quadrimoment total, au sens suivant : la somme vectorielle des quadrimoments des corpuscules émergents i est égale à la somme des quadrimoments des corpuscules incidents a et b :

$$p^{tot} = p^a + p^b = \Sigma\{p^i \mid i \in I\}.$$

L'espace quadridimensionnel est muni d'une forme quadratique de signature mixte (+---) , la métrique de Minkowski.

$$\|\{ct,x,y,z\}\|^2 = c^2t^2 - x^2 - y^2 - z^2 \ ;$$

$$\|\{En/c;P_x , p_y , p_z\}\|^2 = En^2/c^2 - p_x^2 - p_y^2 - p_z^2 \ ;$$

un quadrivecteur v est dit de type *temps* si $\|v\|^2 > 0$; de type *espace* si $\|v\|^2 < 0$; la séparation étant faite par le cône des vecteurs de carré de norme nulle ou cône de *lumière*. Le quadrimoment p^i d'un corpuscule i est un vecteur de type temps, dont la norme est $c\mu^i$, où μ^i désigne la masse du corpuscule i : $\|p^i\|^2 = c^2(\mu^i)^2$: donc l'ensemble des quadrimoments p^i possibles pour un corpuscule i de masse μ^i est une hypersurface H^i de l'espace quadridimensionnel E, ayant pour équation :

$$H^i = \{p | p \in E \ ; \ \|p\|^2 = c^2(\mu^i)^2; \ En > 0\} \ ;$$

H^i est un hyperboloïde ayant le cône de lumière pour asymptote ; plus exactement, H^i est une nappe d'hyperboloïde (à cause de la condition En = énergie positive) : on dira en bref que H^i est la *nappe de masse* du corpuscule i.

3.1 L'espace F des événements : Dans la présente étude, chaque événement est considéré comme un point, dans un espace multidimensionnel F. Comme au § 2, on veut étudier un ensemble Ens de ces points. Jusqu'à présent les études statistiques multidimensionnelles concernent un ensemble d'événements relevant tous de la même réaction, et pour lesquels , de plus, les quadrimoments p_a et p_b des corpuscules incidents sont fixés (avec une certaine précision) par les conditions d'expérience : donc l'ensemble I est fixé , ainsi que $p^{tot} = p^a + p^b = \Sigma\{p^i\}$(Pour une étude multidimensionnelle où le nombre Card I de corpuscules émergents varie avec les événements, cf D. Maïti : Statistique des holographes des réactions entre protons à très haute énergie ; à paraître). Un événement est donc complètement décrit par le système $\{p^i | i \in I\}$ des quadrimoments des particules émergentes ; on note :

$$p(I) = \{p^i | i \in I\} \in \Pi\{H^i | i \in I\} = H^I \subset E^I \ ;$$

un événement est un point p(I) du produit H^I des nappes de masse ; et H^I est inclus dans E^I, produit de Card I = n exemplaires de l'espace quadridimensionnel E.

La distance entre deux événements p(I) et p'(I) est définie par :

$$\|p(I) - p'(I)\|^2 = \Sigma\{-\|p^i - p'^i\|^2 \ i \in I\} = Dis^2(p(I),p'(I)) \ .$$

dans cette formule $\|p^i - p'^i\|^2$ désigne la forme quadratique de Minkowski définie ci-dessus, i.e. :

$$\|p^i - p'^i\|^2 = ((En^i - E'n^i)/c)^2 - (p_x^i - p'^i_x)^2 - (p_y^i - p'^i_y)^2 - (p_z^i - p'^i_z)^2;$$

on peut montrer que, parce que p^i et p'^i sont des quadrivecteurs de même norme $c\mu^i$, appartenant à la nappe $H^i(En > 0)$, $\|p^i - p'^i\|^2$ est négatif ; c'est pourquoi on a introduit un signe - dans la définition de $\|p(I) - p'(I)\|^2$, qui est ainsi une quantité strictement positive , ne s'annulant que si p(I) = p'(I).

Comme espace F, on considère le produit H^I des nappes, muni de la métrique définie ci-dessus. Toutefois une difficulté se présente alors pour le calcul de la moyenne entre deux événements (points de F) . Soit p'(I) et p"(I) deux événements munis des masses respectives m' et m" (masses au sens des pondérations statistiques ; non des masses physiques μ^i des corpuscules) l'événement moyen est p(I) :

$$p(I) = (m'\,p'(I) + m''\,p''(I))/(m+m''); \quad \forall i: p^i = (m'p'^i + m''p''^i)/(m'+m'');$$

or p(I) n'appartient pas à H^I : le quadrimoment p^i calculé comme moyenne pondérée de p'^i et p''^i (dont la norme est $c\mu^i$) a une norme supérieure à $c\mu^i$. On peut songer à corriger cet effet en remplaçant chacun des p^i calculés comme moyenne par un quadrivecteur de norme $c\mu^i$ qui lui est proportionnel ; mais alors p^{tot} n'aurait pas même valeur pour p(I)que pour p'(I) et p"(I). En fait dans la pratique des algorithmes de classification (cf § 3.3) on a trouvé que cette irrégularité ($p(I) \notin H^I$) n'était pas gênante : les p(I) qu'on est amené à calculer comme moyennes s'écartent peu de H^I ; et les carrés de distances entre ces p(I) sont positifs. Mais il importe de signaler qu'on se place non dans $F = H^I$ (produit des nappes) mais dans E^I (produit de n exemplaires de l'espace quadridimensionnel); ou au moins dans C^I (produit de n exemplaires du cône positif des vecteurs de type temps.

3.2 Groupe G des isométries et distance orbitale : Au § 1.2 , on part d'un groupe G d'isométries agissant sur l'espace euclidien E et on définit l'action de G sur les systèmes finis M(I). Ici, E est un espace quadridimensionnel muni de la métrique de Minkowski, dont la signature est mixte : l'analyse du groupe des isométries est le groupe de Lorentz GL, groupe des transformations linéaires respectant la métrique de Minkowski (puis précisément on doit se restreindre au groupe dit *orthochrome* des transformations conservant l'orientation du temps). Le groupe G choisi devra être un sous-groupe de GL. Divers choix sont proposés ci-dessous.

3.2.1 *Le groupe de Lorentz orthochrome GL* : L'orbite GL p(I) d'un événement p(I) est définie par :

$$GL\ p(I) = \{gp(I) \mid g \in GL\} \ ; \quad \text{avec :}$$

$$p(I) = \{p^i \mid i \in I\} \ ; \ gp(I) = \{gp^i \mid i \in I\} \ .$$

parce que GL conserve la norme minkowskienne on a $\|gp^i\| = \|p^i\|$, donc si $p(I) \in H^I$ (produit des nappes, il en est de même de gp(I) : pour trouver la distance orbitale Dor entre deux événements p(I), p'(I) :

$$Dor(p(I),\ p'(I)) = \inf\{Dis(p(I),\ gp'(I)) \mid g \in GL\},$$

on doit appliquer la méthode de Bourgeois, dans le cas d'une forme quadratique de signature quelconque : cela est possible ; mais n'a pas été appliqué effectivement jusqu'ici.

3.2.2 *Le groupe euclidien tridimensionnel GE* : Les événements p(I),p'(I), entre lesquels on aura à calculer la distance orbitale, ont tous un même quadrimoment total p^{tot} : il est donc raisonnable de se restreindre au sous-groupe GE de GL laissant fixe p^{tot}. Autrement dit pour calculer la distance Dor entre une gerbe émergent p(I) et une autre gerbe p'(I), on considère seulement les gp'(I) égaux à p'(I) (au sens de la métrique et du produit scalaire minkowskiens). Comme il est classique, rapportons l'espace quadridimensionnel E à un système de coordonnées dont l'axe des

temps est orienté suivant p^{tot} ; (c'est ce qu'on appelle en physique le système du centre de masses) : on voit que le groupe GE n'est autre que le groupe de transformations laissant fixe l'axe des temps, et agissant sur les trois coordonnées spatiales x,y,z (ou p_x , p_y , p_z) comme une isométrie euclidienne usuelle. Voyons comment se présente maintenant le calcul de Dor. On a :

$$p^i = \{En^i/c\ ;\ p_x^i, p_y^i, p_z^i\} = \{En^i/c; \vec{p}^i\}\ ;\quad gp^i = \{En^i/c; g\vec{p}^i\}:$$

g n'agit que sur les trois composantes spatiales, i.e. sur

$\vec{p} = \{p_x , p_y , p_z\}$. Pour la distance on a :

$$\|p^i - gp'^i\|^2 = (En^i/c - E'^i_n/c)^2 - |\vec{p}^i - g\vec{p}'^i|^2\ ;$$

dans cette formule on a noté entre simples barres la norme euclidienne usuelle :

$$|\vec{p}^i - g\vec{p}'^i|^2 = (p_x^i - gp'^i_x)^2 + (p_y^i - gp'^i_y)^2 + (p_z^i - gp'^i_z)^2.$$

Il apparaît que dans $\|p(I) - gp'(I)\|^2$ seule varie avec g la partie liée aux composantes spatiales : notons :

$$\vec{p}(I) = \{\vec{p}^i \mid i \in I\}\ ;\ |\vec{p}(I) - g\vec{p}'(I)|^2 = \Sigma\{\|\vec{p}^i - gp'^i\|^2 \mid i \in I\}\ ;$$

il vient :

$$Dis(p(I), gp'(I)) = -\Sigma\{(En^i - E'^i_n)/c)^2\} + |\vec{p}(I) - g\vec{p}'(I)|^2.$$

Pour calculer Dor, il suffit de trouver l'isométrie (tridimensionnelle, euclidienne, usuelle) rendant minimum $|\vec{p}(I) - g\vec{p}'(I)|^2$; c'est le problème résolu par Bourgeois (1978). Toutefois les calculs effectués en 1977 (P. Lutz, thèse ; et D. Maïti, thèse) n'ont pas bénéficié de cette solution, dont l'application aux données physique reste donc à faire.

3.2.3 *Le groupe GF des isométries autour de l'axe des faisceaux* : Dans le système du centre de masse ; les moments spatiaux $\vec{p}^a$, $\vec{p}^b$ des particules incidentes sont deux vecteurs opposés définissant un axe appelé l'axe du faisceau. En toute rigueur, pour que deux événements p(I) et p'(I) puissent être considérés comme physiquement identiques (même si les p^i , p'^i diffèrent) il faut qu'il existe une isométrie g telle que p(I) = gp'(I) et que de plus $gp^a = p^a$; $gp^b = p^b$; i.e. l'isométrie g doit respecter l'axe du faisceau (en effet l'événement dans sa totalité comprend la gerbe émergente avec les deux particules incidentes). Il est donc justifié de se restreindre à un sous-groupe euclidien GE : le groupe GF des isométries laissant fixes p^a , p^b ; i.e. le groupe des rotations-symétries autour de l'axe du faisceau. Dès lors on est exactement dans le cas complètement résolu par le *problème* (au § 2.6 : calcul de θ et ε).

3.2.4 *Le groupe GA des isométries autour de l'axe principal* : On sait (cf D. Maïti : thèse ; D. Maïti & M.C. Touboul, à paraître ···)
que pour chaque événement le système $\vec{p}(I)$ des vecteurs $\vec{p}^i$ (considérés dans le système du centre de masse où leur somme p^{tot} est nulle) peut être ajusté (selon le critère des moindres carrés ; par analyse factorielle) à une droite A (le premier axe principal ; appelé ici en bref l'axe principal) qui ne coïncide pas exactement avec l'axe du faisceau (même si elle en diffère très peu dans le cas des réactions à haute énergie , où

la longitudinalité est très marquée : cf [Phys.Cor.] II, § 4.2.2 *in Cahiers* Vol II p 463) : pour comparer deux gerbes émergentes p(I), p'(I), il est intéressant d'amener d'abord celles-ci par déplacement à avoir même axe principal A, puis de déterminer la distance Dor par rotation - symétrie autour de A.

3.3 *Les classifications effectuées* : On a appliqué l'algorithme de CAH et l'algorithme de E. Diday, modifiés pour utiliser la distance orbitale (§§ 2.3 & 2.4) ; celle-ci étant calculée soit pour le groupe GF des isométries autour de l'axe du faisceau (§ 3.2.3 : P. Lutz, thèse), soit pour le groupe GA des isométries autour de l'axe principal (§ 3.2.4 : D. Maïti, thèse). Les résultats publiés en détail dans les thèses concernent la réaction :

$$K^- p \to K^-, p, \pi^+, \pi^- \text{ à } 14{,}3 \text{ GeV/c.}$$

(i.e. faisceau de K^- à 14,3 GeV/c, contre cible de protons). Environ 17.000 événements étaient disponibles ; on les a d'abord agrégés en quelques centaines de classes (ou îlots) par la recherche des plus proches voisins : cette étape effectuée avant les recherches rapportées ici a demandé des heures de calcul : on la referait tout autrement aujourd'hui. Les classifications publiées dans les thèses portent donc sur enviton 400 îlots. En étudiant les classes obtenues, on aboutit aux conclusions suivantes que nous soumettrons seules ici aux lecteurs statisticiens :

1 une classe peut correspondre à plusieurs sous-canaux (pour l'explication de ce terme, cf [Phys. Cor.] § 3.5 : Vol II, n° 2, p 327).

2 il existe des interférences entre les sous-canaux

3 les classes qui interfèrent entre elles ne peuvent être séparées qu'à un niveau très bas de la hiérarchie.

Après les recherches objets des thèses, on a appliqué la méthode à d'autres données : un sommaire des résultats obtenus se trouve dans P. Lutz & D. Maïti, *in Actes du 3e Topical Meeting on Multidimensional Analysis of High-Energy Data, Nijmegen* 1978.

Quant au temps de calcul, malgré la complexité relative de la formule de distance orbitale, il est devenu praticable même en CAH grâce à l'algorithme des voisinages réductibles de M. Bruynooghe (cf Cahiers ; Vol III pp 7-33 ; 1978) : e.g. 10 min de temps C.P. sur CDC 6600, pour classer 1000 événements.