ACP et recherche de photons avec le calorimètre électromagnétique de l'expérience PHENIX

13 juin 2003

0.1 Motivations

L'utilisation des photons directs pour la mise en évidence du Plasma de Quark et de Gluons (PQG) (mettre une référence ici) est un objectif essentiel de l'expérience PHENIX (mettre une référence ici). Or, la détection des photons directs s'avère relativement difficile : ils sont noyés au milieu de photons de décroissance produits majoritairement par les nombreux π^0 issus des collisions. Précisons que la sélection des photons s'effectue habituellement par une série de coupures (temps de vol inférieur à une valeur donnée, χ^2 sur la forme attendue de la gerbe inférieur à une certaine valeur, dispersion de la gerbe inférieure à une autre valeur, etc.). Mais l'identification des photons par cette technique s'est révélée difficile, notamment en raison de performances relativement moins idéales que celles attendues sur le temps de vol lors de la première année de prise de données (mettre une référence ici).

Ceci nous conduisit donc à envisager un meilleur moyen d'identifier les photons. L'idée fut lancée d'utiliser l'Analyse en Composantes Principales (ACP). Cette technique d'analyse permet de trouver les axes caractéristiques d'une distribution dans un espace comportant un grand nombre de dimensions, du moins trop grand pour pouvoir être visualisé.

0.2 Méthode générale de l'ACP

L'ACP est une méthode de base de l'analyse de données. Elle consiste à étudier un ensemble de n *individus* possédant p *caractères* numériques.

0.2.1 Présentation succincte

L'étude de deux caractères x^1 et x^2 ne présente en général aucune difficulté étant donné qu'il est facile d'observer leur corrélation. Celle-ci apparaît immédiatement dans un graphique plan à deux dimensions. Ce graphique contient tous les individus e_i, représentés par leurs coordonnées x_i^1 et x_i^2. Grâce à ce graphique, la liaison entre x^1 et x^2 et les groupes d'individus sont faciles à repérer (cf. les exemples de la Fig. 1).

Avec trois caractères, la situation devient plus difficile à analyser et même impossible à partir de quatre. Mais supposons que l'on veuille tout de même représenter nos individus dans un plan. Il faudrait alors exécuter une projection, ce qui aura certainement pour effet de déformer fortement la distribution des individus : la distribution dans le plan ne sera en rien représentative de ce qu'elle était dans l'espace à p dimensions. A moins, bien sûr, qu'il n'existe n-2 relations linéaires liant les caractères entre eux.

Le but de l'ACP va être de rendre les distortions après projection les plus petites possibles. Pour cela, il faut faire un peu de géométrie. Nos individus

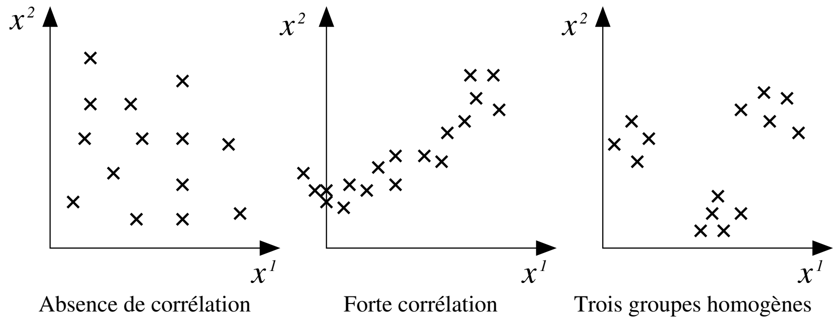

FIG. 1 – quelques exemples de distributions à deux dimensions.

$\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_n$ sont projetés dans un plan définis par deux droites Δ_1 et Δ_2. Appelons les projetés dans ce plan $\mathbf{f}_i$.

(insérer ici un dessin représentant la situation)

Ce plan de projection devra être choisi de manière à conserver au mieux les distances. Or, la projection raccourcit les distances : $d(\mathbf{f}_i\,;\mathbf{f}_j) \leq d(\mathbf{e}_i\,;\mathbf{e}_j)$. L'un des objectifs de l'ACP va donc être de maximiser la distance entre les $\mathbf{f}_i$.

Pour déterminer ce plan de projection, appelé *plan principal*, il suffit de trouver deux droites Δ_1 et Δ_2. Si ces deux droites sont choisies perpendiculaires, la relation suivante est valide :

$$d^2(\mathbf{f}_i\,;\mathbf{f}_j) \leq d^2(\boldsymbol{\alpha}_i\,;\boldsymbol{\alpha}_j) + d^2(\boldsymbol{\beta}_i\,;\boldsymbol{\beta}_j)$$

où les $\alpha_i$et les β_i sont les projections des $\mathbf{e}_{\mathrm{i}}$ (et a fortiori, des $\mathbf{f}_i$) sur Δ_1 et Δ_2 respectivement. L'indépendance de nos deux droites est une aide considérable. La moyenne des carrés des distances des $\mathbf{f}_i$ valant simplement la somme des moyennes des carrés des distances des α_i et des β_i.

Toute la méthode consiste alors à trouver $\Delta_1$maximisant la moyenne des $d^2(\boldsymbol{\alpha}_i\,;\boldsymbol{\alpha}_j)$, puis à trouver Δ_2 normale à Δ_1 maximisant la moyenne des $d^2(\boldsymbol{\beta}_i\,;\boldsymbol{\beta}_j)$. Cette démarche si elle est menée à terme conduit à trouver les droites Δ_3, ..., Δ_p, toutes perpendiculaires entre elles. Ces droites Δ_k sont appelées *axes principaux* du nuage.

De nouvelles coordonnées $(c_i^1, c_i^2, ..., c_i^p)$ sont obtenues en projetant $\mathbf{e}_i$ de coordonnées initiales $(x_i^1, x_i^2, ..., x_i^p)$ sur les axes principaux. Cela revient implicitement à construire p nouveaux caractères $\mathbf{c}^1$, ..., $\mathbf{c}^p$. Ces nouveaux caractères $\mathbf{c}^k$ sont appelés *composantes principales*. Ces caractères, qui ne sont en fin de compte que la liste des n individus sur l'axe principal Δ_k, sont en fait une combinaison linéaire des caractères initiaux :

$$\mathbf{c}^k = u_1^k\mathbf{x}^1 + u_2^k\mathbf{x}^2 + ... + u_p^k\mathbf{x}^p$$

Les coefficients $(u_i^1, u_i^2, ..., u_i^p)$ sont les composantes du k-ième facteur principal.

Nous disposons donc maintenant des p composantes principales. Le choix de q caractères ($q < p$) choisis de manière à diminuer les déformations de la distribution initiale dans l'espace des caractères à p dimensions sera de choisir parmi les premières composantes principales, en rappelant que celles-ci sont ordonnées. Cette réduction du nombre de caractères pertinents se fonde sur une représentation géométrique ; elle est impossible si les p caractères initiaux ne sont pas *linéairement dépendants*. Les coefficients de corrélation des caractères doivent donc être non-nul pour qu'il soit possible d'appliquer l'ACP.

L'ACP se présente donc comme une méthode *factorielle* et *linéaire* de réduction des caractères : *factorielle* car la réduction ne s'effectue pas par l'élimination de certains caractères mais par la formation des nouveaux caractères synthétiques au moyens de *facteurs*, et *linéaire* car il s'agit ici de ne travailler qu'avec des *combinaisons linéaires*. Notons encore que l'ACP ne fait pas de différence entre les individus, ceux-ci étant totalement identiques quant à leur prise en compte au moment de l'analyse même.

L'utilisation de combinaisons linéaires nous conduit a envisager un traitement matriciel de l'ACP. Nous pouvons en effet construire deux espaces vectoriels à p et n dimensions avec nos données, et les transformations peuvent être exécutées simplement à l'aide de matrices.

0.2.2 Résumés numériques

Les p caractères numériques x^j sont résumables indépendamment les uns des autres par leurs moyennes $\overline{x^j}$ et leurs écarts types s^j. L'individu fictif dont les coordonnées sont les valeurs moyennes de tous les caractères s'appelle le centre de gravité du nuage, noté $\mathbf{g}$:

$$\mathbf{g} = (\overline{x^1}, \overline{x^2}, ..., \overline{x^p})$$

Pour matérialiser les liaisons entre les p caractères pris deux à deux, il faut calculer leurs covariances s_{jk} ou leurs coefficients de corrélation r_{jk}. Cela fait $\frac{p(p-1)}{2}$ coefficients à calculer. L'ensemble des variances et des covariances forme la matrice de variance V des p caractères, où les termes diagonaux s_{jj} sont en fait les variances s_j^2. Notons d'emblée que V est une matrice carrée symétrique (car $s_{jk} = s_{kj}$) d'ordre p.

$$\mathrm{V} = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ & s_2^2 & & \vdots \\ & & \ddots & \\ & & & s_p^2 \end{pmatrix}$$

De même que nous pouvons écrire la matrice de variance, nous pouvons identiquement écrire la matrice de corrélation R constituée des coefficients de corrélation entre les p caractères. Cette matrice est elle aussi carrée d'ordre p et symétrique, les termes diagonaux valant ici 1 (ce qui est logique, puisque un caractère donné est totalement corrélé avec lui-même).

$$\mathrm{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ & 1 & & \vdots \\ & & \ddots & \\ & & & 1 \end{pmatrix}$$

La relation qui unie R et V s'obtient en utilisant la matrice diagonale suivante :

$$\mathrm{D}_{1/s} = \begin{pmatrix} 1/s_1 & & & \bigcirc \\ & 1/s_2 & & \\ & & \ddots & \\ \bigcirc & & & 1/s_p \end{pmatrix}$$

on a alors la relation suivante :

$$\mathrm{R} = \mathrm{D}_{1/s} \mathrm{V}\, \mathrm{D}_{1/s}$$

Il est possible de calculer V directement à partir des données initiales. Il faut cependant que les individus soient *centrés*, c'est-à-dire que les moyennes des différents caractères doivent alors être nulles. Ceci revient en fait à placer l'origine des axes sur le centre de gravité **g** du nuage, l'opération effectuées pur y parvenir étant alors une simple translation. Appelons X la matrice à n lignes et p colonnes contenant les données centrées, nous obtenons alors la relation suivante :

$$\mathrm{V} =^t \mathrm{XDX}$$

où D est la matrice des poids (d'ordre n) :

$$\mathrm{D} = \begin{pmatrix} p_1 & & & \bigcirc \\ & p_2 & & \\ & & \ddots & \\ \bigcirc & & & p_n \end{pmatrix}$$

dans le cas le plus général, aucun individu n'est privilégié et tous les points valent $\frac{1}{n}$. Pour la suite, la supposition que les caractères sont centrés sera implicite.

0.2.3 L'espace des individus

L'espace des individus est un espace vectoriel $\mathbb{R}^p$. Chaque individu e_i y est représenté par un vecteur à p dimensions $\mathbf{e}_i$ de coordonnées $(x_i^1, x_i^2, ..., x_i^p)$.

0.2.3.1 La métrique

Dans le cadre de l'ACP, il faut calculer des distances. Or, la métrique prend une importance capitale ici, d'autant plus que les différents axes possèdent différentes unités. En physique, la distance est donnée par la formule de Pythagore : le carré de la distance vaut la somme des carrés des différences des coordonnées (cf. Fig. 2). Il existe toutefois un préalable, il faut impérativement que les coordonnées soient toutes exprimées dans la même unité.

$d^2 = (x_1^k - x_2^k)^2 + (x_1^j - x_2^k)^2$

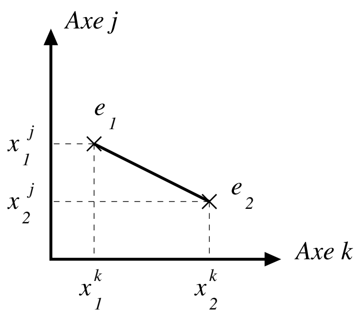

FIG. 2 – distance entre deux individus e_1 et e_2

Cependant, en statistique, les caractères ne sont pas exprimés dans les mêmes unités. Il est alors difficile de calculer une distance viable, surtout si les caractères possèdent des unités incompatibles entre elles. La formule de Pythagore n'est alors plus une définition convenable de la distance, elle prendrait un caractère aussi arbitraire qu'une autre définition. D'autre part, l'adoption d'axes orthogonaux en statistique est purement arbitraire. Des axes formant un angle quelconque θ auraient tout aussi bien pu convenir, comme illustré par la figure 3 :

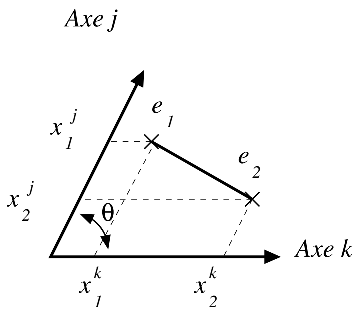

FIG. 3 – distance entre deux individus e_1 et e_2 dans le cas d'axes obliques

La distance entre les deux individus devient alors :

$$d^2 = (x_1^k - x_2^k)^2 + (x_1^j - x_2^j)^2 - 2(x_1^k - x_2^k)(x_1^j - x_2^j)\cos\theta$$

respectant ainsi le produit scalaire dans un cas ou les axes ne sont pas orthogonaux. De manière générale, la distance entre deux individus s'écrit :

$$d^2(e_1\,;e_2) = \sum_{k=1}^{p}\sum_{j=1}^{p}(x_1^k - x_2^k)(x_1^j - x_2^j)$$

ou si l'on note ceci sous forme matricielle :

$$d^2(e_1\,;e_2) =^t (\mathbf{e}_1 - \mathbf{e}_2)\mathrm{M}(\mathbf{e}_1 - \mathbf{e}_2)$$

M est une matrice quelconque définie positive, et s'appelle la *métrique* de l'espace. Le produit scalaire entre deux vecteurs d'un espace doté d'une métrique M s'écrit alors :

$$< \mathbf{e}_1\,;\mathbf{e}_2 >_{\mathrm{M}} =^t \mathbf{e}_1\mathrm{M}\mathbf{e}_2$$

Dans le cas d'un espace euclidien, la métrique M est alors la matrice unité I. $\|\mathbf{e}_1\|_{\mathrm{M}}^2$ est le produit scalaire de $\mathbf{e}_1$ par lui-même, et $\|\mathbf{e}_1\|_{\mathrm{M}}$ est sa M-norme, c'est à dire sa longueur.

Dans l'optique de l'ACP, les métriques les plus utilisées sont les métriques diagonales. Leur emploi équivaut à pondérer les caractères. La métrique la plus utilisée est la métrique D_{1/s^2} :

$$\mathrm{M} = \mathrm{D}_{1/s^2} = \begin{pmatrix} 1/s_1^2 & & & \bigcirc \\ & 1/s_2^2 & & \\ & & \ddots & \\ \bigcirc & & & 1/s_p^2 \end{pmatrix}$$

ce qui revient à diviser les caractères par leur propre écart type. De ce fait, les différences d'unités de mesure entre les caractères disparaissent car x_j/s_j n'a pas de dimension, et la distance devient alors valide au sens physique du terme. Par ailleurs, tout changement d'unité sur un caractère sera effectif aussi sur son écart type : les unités n'ont plus d'importance grâce à la métrique D_{1/s^2}. Autre avantage de cette métrique, il ne sera plus possible de donner trop d'importance à un caractère parce que son écart type est bien plus grand que celui des autres caractères.

Utiliser une matrice diagonale en tant que métrique revient à multiplier chaque caractère par un coefficient précis. Par exemple, se servir de la métrique suivante $\mathrm{D}_a = \begin{pmatrix} a_1 & & & \bigcirc \\ & a_2 & & \\ & & \ddots & \\ \bigcirc & & & a_p \end{pmatrix}$ revient à multiplier les caractères par $\sqrt{a_i}$, et à utiliser ensuite la matrice unité I comme métrique. Plus généralement, pour toute matrice M symétrique définie positive M, il existe une matrice T (en réalité, il en existe une infinité) telle que $M =^t TT$. Le produit scalaire peut alors être ré-écrit :

$$\begin{aligned}\langle \mathbf{e}_1 ; \mathbf{e}_2 \rangle_M &= {}^t\mathbf{e}_1 M \mathbf{e}_2 \\ &= {}^t\mathbf{e}_1 {}^t T T \mathbf{e}_2 \\ &= {}^t(T\mathbf{e}_1)(T\mathbf{e}_2) \\ &= \langle T\mathbf{e}_1 ; T\mathbf{e}_2 \rangle_I\end{aligned}$$

donc le produit scalaire peut être obtenu en transformant les données avec la matrice T puis en calculant le produit scalaire dans un espace euclidien. Cette gymnastique revient à remplacer le tableau de données X par $Y = X^tT$ et à prendre pour métrique la matrice unité I.

0.2.3.2 Coordonnées des individus sur un nouvel axe

Par convention, le référentiel utilisé pour représenter l'espace des individus est orthonormé. Les individus sont représentés par leurs coordonnées sur les axes des caractères x^1, x^2, ...,x^p.

Soit une droite quelconque Δ. La projection des individus sur cette droite donne naissance à un nouveau caractère $\mathbf{c}$ dont les valeurs c_1, c_2, ..., c_n sont les valeurs algébriques des projections des individus $\mathbf{e}_i$ sur cette droite (cf. figure 4).

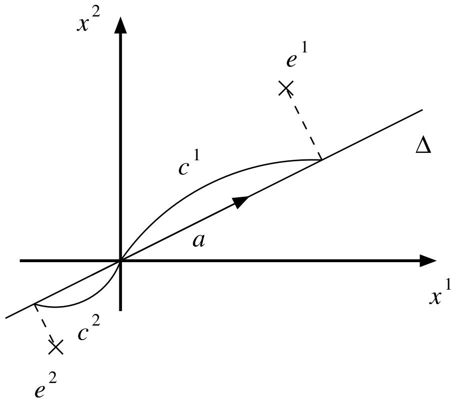

FIG. 4 – caractère synthétique $\mathbf{c}$ dans l'espace des individus

Appelons $\mathbf{a}$ le vecteur unitaire de Δ, de M-norme I. La valeur de c_1, résultat de la projection orthogonale de l'individu $\mathbf{e}_1$ sur Δ, est égale au produit scalaire de $\mathbf{e}_1$ par $\mathbf{a}$: $c_1 =^t \mathbf{e}_1 M\mathbf{a} =^t (M\mathbf{a})\mathbf{e}_1$ car M est symétrique ; soit $\mathbf{u} = M\mathbf{a}$, alors $c_1 =^t \mathbf{u}\mathbf{e}_1 = \sum_{j=1}^{p} u_j x_1^j$. Le caractère $\mathbf{c}$ dont les valeurs sont les n coordonnées c_1, c_2, ..., c_n s'obtient simplement par la formule $\mathbf{c} = X\mathbf{u}$. Nous retrouvons ici le caractère factoriel de l'ACP : **c** est une combinaison linéaire des p caractères initiaux au moyen du facteur **u**. D'autre part, si $\mathrm{M} = \mathrm{I}$, il y a égalité entre le facteur **u** et le vecteur unitaire **a** de Δ.

Autre propriété intéressante, si l'axe Δ passe par l'origine, le caractère **c** sera alors centré, car l'origine est confondue avec le centre de gravité du nuage.

0.2.3.3 L'inertie du nuage

L'inertie du nuage de points est la moyenne des carrés des distances des n points au centre de gravité, c'est-à-dire l'origine dans notre repère centré :

$$\mathcal{I} = \sum_i p_i \left\|\mathbf{e}_i\right\|_{\mathrm{M}}^2 = \sum_i p_i \, {}^t\mathbf{e}_i \mathrm{M} \mathbf{e}_i$$

L'inertie est une quantité intrinsèque du nuage. Elle donne une mesure de l'éloignement des points au centre de gravité, donc une dispersion globale du nuage.

Il est possible de montrer que l'inertie est égale à la moyenne des carrés des $\frac{n(n-1)}{2}$ distances entre les points du nuage. Dans ce cas de figure, le plan principal du nuage peut être interprété comme le plan qui rend l'inertie des n points projetés sur lui maximale. L'inertie peut aussi être définie par rapport à un point quelconque **o** différent du centre de gravité :

$$\mathcal{I}_h = \sum_i p_i d^2(\mathbf{e}_i \,; \mathbf{h})$$

$\mathcal{I}_h$ est reliée à $\mathcal{I}$ par la formule de Huyghens :

$$\mathcal{I}_h = \mathcal{I} + d^2(\mathbf{g} \,; \mathbf{h})$$

$\mathcal{I}_h$ est toujours supérieures à $\mathcal{I}$, la valeur minimale étant obtenue lorsque $\mathbf{h} = \mathbf{g}$. La déduction logique de ceci est que la recherche d'un plan rendant maximale l'inertie des projections des n points est équivalente à la recherche du plan passant « au plus près » de l'ensemble des points du nuage au sens où la moyenne des carrés des distances des points du nuage au plan est minimale.

Soit **h** la projection de **g** sur le plan principal. **h** est donc le centre de gravité des projections des points du nuage sur le plan principal. Soit $\mathbf{f}_i$ le projection de $\mathbf{e}_i$ sur le plan. Le triangle $\mathbf{e}_i \,; \mathbf{f}_i \,; \mathbf{h}$ est rectangle en $\mathbf{f}_i$ (cf. figure 5), d'où : $d^2(\mathbf{e}_i \,; \mathbf{f}_i) = d^2(\mathbf{e}_i \,; \mathbf{h}) - d^2(\mathbf{f}_i \,; \mathbf{h})$ et $\sum_i p_i d^2(\mathbf{e}_i \,; \mathbf{f}_i) = \mathcal{I}_h - \sum_i p_i d^2(\mathbf{f}_i \,; \mathbf{h})$. Or, d'après la formule de Huygens, nous savons que $\mathcal{I}_h = \mathcal{I} + d^2(\mathbf{g} \,; \mathbf{h})$. Il devient clair que pour rendre minimale la moyenne des carrés des distances entre les individus $\mathbf{e}_i$ et leurs projetés $\mathbf{f}_i$, il faut nécessairement que $\mathbf{g} = \mathbf{h}$ et que l'inertie du nuage projetée $\sum_i p_i d^2(\mathbf{f}_i \,; \mathbf{h})$ soit maximale. Nous savons donc que le plan

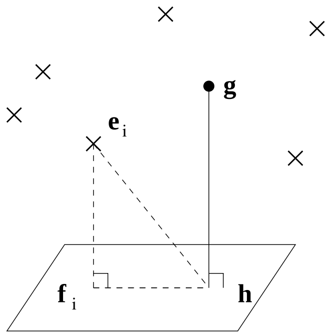

FIG. 5 – recherche du plan principal grâce à l'inertie

principal passe par le centre de gravité du nuage, cette propriété pouvant être étendue aux axes principaux. Par la suite, cette propriété des axes principaux et du plan principales sera toujours implicite.

Il est possible de démontrer que $\mathcal{I}$ s'exprime simplement en fonction de la métrique M et de la matrice de variance V :

$$\mathcal{I} = \text{Trace}\,(\text{MV})$$

D'où les déductions suivantes :

- si $\mathbf{M} = \mathbf{I}$ l'inertie est la somme des variances des p caractères
- si $\mathbf{M} = \mathbf{D}_{1/s^2}$:

$$\begin{aligned}\text{Trace}\,(\text{MV}) &= \text{Trace}\,(\text{D}_{1/s^2}\text{V}) = \text{Trace}\,(\text{D}_{1/s}\text{VD}_{1/s}) \\ &= \text{Trace}\,\text{R} = p\end{aligned}$$

  l'inertie vaut simplement le nombre de caractères
- si M est quelconque, nous pouvons toujours utiliser la relation vue précédemment $\text{M} =^t \text{TT}$, et dans ce cas nous avons :

$$\begin{aligned}\text{Trace}\,(\text{MV}) &= \text{Trace}\,(^t\text{TT}^t\text{X}^t\text{DX}) = \text{Trace}\,(\text{T}^t\text{XDX}^t\text{T}) \\ &= \text{Trace}\,(^t\text{YDY})\end{aligned}$$

0.2.4 L'espace des caractères

Chaque caractère x^j est défini par les valeurs que prennent les n individus. Les caractères seront donc représentés comme des vecteurs $\mathbf{x}^j$ à n dimensions, et l'espace des caractères sera donc un espace à n dimensions noté $\mathbb{R}^n$

0.2.4.1 La métrique de l'espace des caractères

Encore une fois, comme dans l'espace des individus, il faut mesurer des distances entre les caractères, et pour cela il faut tout d'abord définir une métrique pour l'espace des caractères. Rappelons que la métrique est représentée par une matrice symétrique définie positive. Contrairement à ce qui se passe dans l'espace des individus, le choix de la métrique de l'espace des caractères est presque évident : la matrice des poids D s'impose. Le produit scalaire dans ce cas s'écrit :

$$\begin{aligned}\left\langle \mathbf{x}^j ; \mathbf{x}^k \right\rangle_{\mathrm{D}} &= {}^t\mathbf{x}^j \mathrm{D} \mathbf{x}^k \\ &= \sum_{i=1}^{n} p_i x_i^j x_i^k \\ &= s_{jk}\end{aligned}$$

qui n'est autre que la covariance entre ces deux caractères (n'oublions pas que les caractères sont centrés et que leur moyenne est ainsi nulle). Et par voie de conséquence, il est aisé de voir que $\|\mathbf{x}^j\|^2 = s_j^2$, et que la norme d'un caractère est égale à son écart type.

D'autre part, l'angle θ entre deux vecteurs dans un espace euclidien est défini à partir du produit scalaire de la façon suivante :

$$\cos\theta_{jk} = \frac{\left\langle \mathbf{x}^j ; \mathbf{x}^k \right\rangle}{\|\mathbf{x}^j\| \, \|\mathbf{x}^k\|} = \frac{s_{jk}}{s_j s_k} = r(j ; k)$$

et nous retrouvons la corrélation linéaire entre les deux caractères. Les angles entre caractères sont donc très simples à étudier avec la métrique D ce qui justifie pleinement le choix de D comme métrique de l'espace des caractères.