LES CAHIERS DE L'ANALYSE DES DONNÉES

J.-P. BENZÉCRI

Problèmes statistiques et méthodes géométriques

Les cahiers de l'analyse des données, tome 3, n° 2 (1978), p. 131-146

<http://www.numdam.org/item?id=CAD_1978__3_2_131_0>

Les Cahiers de l'Analyse des Données
Vol. III - 1978 - n° 2 - p. 131-146

PROBLÈMES STATISTIQUES ET MÉTHODES GÉOMÉTRIQUES

[RÉGR. GÉOM.]

par J.-P. Benzécri (1)

Le présent article écrit en 1970 expose après le problème général de la régression les principes de deux méthodes conçues pour se prémunir contre des résultats illusoires car instables. Ces méthodes ont été développées et appliquées par P. Cazes qui les a de plus comparées à d'autres méthodes, notamment à la régression par l'analyse factorielle des correspondances. Nous publions également dans ce cahier le premier article d'une série donnant un exposé d'ensemble de la régression, article dû à P. Cazes et relatif à la régression sous contrainte, et publierons ultérieurement outre la suite de la série, des études d'applications particulières.

L'énoncé le plus général du problème de la régression nous paraît être ceci : en fonction de ce que l'on sait, donner une expression approchée de ce que l'on ignore. Pour certains, c'est là le thème de toute la science, qui n'aurait pour but que de prévoir. Assurément, la capacité de prévoir est le signe qu'une discipline a atteint un objet réel : c'est par ce critère que l'on réduit à néant tant de prétentieuses divagations. Mais la fin de la science est autre que de satisfaire à ce critère : elle est de connaître, non de prévoir. Par delà la distinction temporelle entre ce que l'on rencontre d'abord, et ce qu'on prévoit qu'on trouvera ensuite, l'esprit aspire à découvrir quelque chose de l'ordre même du tout. Voilà pourquoi, dans la méthodologie statistique nous donnons la place d'honneur à l'analyse globale des données. Cependant nous traiterons ici de la régression de l'ajustement et de l'estimation, qui jouent parfois un rôle utile.

D'un énoncé général, il est rare qu'on passe immédiatement à une technique praticable. Il nous faut délimiter pour la régression un domaine qui laisse prise au calcul. Ce domaine mathématique sera, lui, trop étroit pour contenir un objet réel : mais nous en tirerons des règles (§§ 3 & 4)) qui, prudemment appliquées, contribueront au programme indéfini proposé d'abord. Rappelons donc le modèle probabiliste (§ 1) avant d'en critiquer la portée réelle (§ 2).

1 *Le modèle probabiliste* : Soit Ω un espace probabilisé, c'est-à-dire, en bref, un ensemble sur lequel sont définies les notions de partie mesurable (ou fonction pour laquelle l'image réciproque de tout intervalle de R est une partie mesurable de Ω), et d'élément de volume positif $d\omega$, (ou élément de probabilité), la masse totale de Ω étant 1 : $\int_\Omega d\omega = 1$. Du point de vue des calculs, on pourra sans erreur interpréter les formules qui suivent comme si Ω était un cube d'arête unité, (ou même l'intervalle (0,1)) muni de l'élément de volume usuel. Du point de vue des applications on se référera au cas où Ω est l'ensemble de toutes les formes possibles de crânes de rongeurs ; une forme ω étant parfaitement caractérisée par, disons, 100 mensurations, Ω sera un domaine de R^{100}, et la mesure d'une partie A de Ω sera la probabilité que la forme ω d'un crâne tombe dans A.

Un point ω de Ω est appelé événement. Une fonction mesurable (ici, il s'agira de fonctions à valeurs réelles) sur Ω est appelée fonction aléatoire, ou parfois variable aléatoire : $f(\omega)$ est la valeur pour l'événement ω de la variable aléatoire f. (Dans l'exemple des crânes, f pourra être le volume intérieur, ou le périmètre etc.). On note L_2^Ω (ou, en bref, L_2) l'ensemble

(1) *Professeur de statistique. Université Pierre et Marie Curie. Paris.*

des fonctions aléatoires f pour lesquelles l'intégrale $\int_\Omega |f(\omega)|^2 \, d\omega$ est bornée ; L_2 est un espace de Hilbert, où est défini le produit scalaire: $\langle f,g \rangle = \int_\Omega f(\omega)\, g(\omega)\, d\omega$. Un élément f de L_2 est appelé communément en analyse une fonction de carré sommable ; ici on dira plutôt : variable aléatoire admettant un moment d'ordre 2 fini :

$$M_t^2(f) = \int_\Omega |f(\omega)|^2 \, d\omega = \|f\|^2 < \infty .$$

Si $f \in L_2$, f admet une moyenne, appelée ici espérance mathématique:

$$\mathrm{Moy}(f) = f = \int_\Omega f(\omega)\, d\omega = \langle 1, f \rangle .$$

(où on a noté 1 la fonction constante de valeur 1). On appelle variance de f le moment d'ordre 2 de f diminuée de sa moyenne :

$$\mathrm{Var}(f) = \int_\Omega |f(\omega) - \mathrm{Moy}(f)|^2 \, d\omega = M_t^2(f) - |\mathrm{Moy}(f)|^2 ;$$

La variance Var(f) mesure la dispersion de f autour de sa moyenne; Var(f) ne s'annule que si f est presque sûrement égale à sa moyenne Moy(f), i.e. si :

$$\mathrm{Mes}\{\omega \mid \omega \in \Omega \ ; \ f(\omega) \neq \mathrm{Moy}(f)\} = 0 ;$$

en ce cas, dans L_2 (qui est, on le sait, en toute rigueur non un espace de fonctions mais un espace de classes de fonctions égales presque partout) f ne se distingue pas de la fonction constante Moy(f). On appelle parfois covariance le produit scalaire :

$$\langle f - \mathrm{Moy}\, f, g - \mathrm{Moy}\, g \rangle = \int_\Omega (f(\omega) - \mathrm{Moy}(f))(g(\omega) - \mathrm{Moy}(g))\, d\omega = \mathrm{Cov}(f,g)$$

Quant au coefficient de corrélation, c'est dans l'espace de Hilbert un cosinus : le cosinus de l'angle formé par les deux vecteurs f - Moy f et g - Moy g :

$$\mathrm{Corr}(f,g) = \mathrm{Cov}(f,g)/(\mathrm{Var}(f)\ \mathrm{Var}(g))^{1/2} .$$

Ainsi pour l'analyse du hasard, les notions géométriques familières reçoivent des noms nouveaux.

A toute variable aléatoire f est associée une mesure positive de masse totale 1 sur R, appelée loi de probabilité de f. Soit A une partie mesurable de R, on a :

$$\mathrm{Prob}(f(\omega) \in A) = \int \{d\omega \mid \omega \in \Omega \ ; \ f(\omega) \in A\}.$$

(où on a noté : $\int \{d\omega \mid \omega \in X\}$ pour $\int_X d\omega$, ce qui évite les superpositions d'indices, et est conforme à l'usage adopté pour les sommes finies). Comme le poids d'une partie A de R pour la loi de f n'est autre que le poids pour $d\omega$ de l'image réciproque de A par f^{-1}, il semble permis de noter ici df^{-1} la mesure sur R qu'est cette loi de f. Plus généralement soit $f^1, f^2, \ldots, f^n$ n variables aléatoires (i.e. n fonctions mesurables sur Ω à valeur réelle). On définit une mesure positive de masse totale 1 sur R^n appelée loi conjointe des n variables aléatoires $\{f^1, \ldots, f^n\}$. Soit A une partie mesurable de R^n ; on a :

$$\mathrm{Prob}(\{f^1(\omega), \ldots, f^n(\omega)\} \in A) =$$
$$\int \{d\omega \mid \omega \in \Omega \ ; \ \{f^1(\omega), \ldots, f^n(\omega)\} \in A\}.$$

De même que dans le cas d'une variable, on pourra noter $d(f^1 x \ldots x f^n)^{-1}$ la mesure sur R^n loi conjointe des n variables aléatoires f^i. (On notera que rien n'impose ici que les f^i soient deux à deux distinctes : si, par exemple, $n = 2$, $f^1 = f^2 = f$, la mesure $d(f \times f)^{-1}$ aura pour support la diagonale de R^2). Supposons que l'application $(f^1 x \ldots x f^n)$ de Ω dans R^n soit injective (ou du moins que soit injective la restriction à une partie de Ω dont la masse est 1) : on pourra alors

identifier Ω avec son image dans R^n par $(f^1 x \ldots x f^n)$, et $d\omega$ avec $d(f^1 x \ldots x f^n)^{-1}$; un événement ω étant caractérisé par la valeur des n grandeurs réelles $f^1(\omega) \ldots f^n(\omega)$ (ce qu'on a fait ci-dessus dans l'exemple d'une forme de crâne décrite par 100 mensurations).

Soit F une fonction de n variables réelles ; et soit n variables aléatoires $f^1, \ldots, f^n$; on a :

$$\int_\Omega F(f^1(\omega), \ldots, f^n(\omega))\, d\omega = \int_{R^n} F(x^1, \ldots, x^n)\, d(f^1 x \ldots x f^n)^{-1}$$

l'espérance mathématique de F peut être calculée dans R^n d'après la loi conjointe des f^i.

Ceci posé, deux variables aléatoires f et g sont dites indépendantes en probabilité (ou stochastiquement indépendantes) si la loi conjointe de {f,g} est le produit des lois de f et de g : ce qu'on écrira $d(fg)^{-1} = df^{-1} \times dg^{-1}$. La condition d'indépendance peut encore s'énoncer directement sans recourir explicitement à la notion de loi : f et g sont indépendantes si quels que soient les intervalles A et B de R on a :

$$\int \{d\omega \mid \omega \in \Omega \; ; \; f(\omega) \in A \; ; \; g(\omega) \in B\} = \int \{d\omega \mid \omega \in \Omega \; ; \; f(\omega) \in A\} \times \int \{d\omega \mid \omega \in \Omega \; ; \; g(\omega) \in B\}.$$

On voit qu'en particulier la constante 1 (ou toute autre constante) forme avec toute variable aléatoire f, une paire de variables indépendantes.

A la notion d'indépendance s'oppose celle de liaison : deux variables qui ne sont pas indépendantes sont liées en ce sens que la conaissance de la valeur de l'une pour un événement ω apporte au moins quelques présomptions nouvelles sur la valeur de l'autre. La forme la plus extrême de liaison est la dépendance fonctionnelle que nous considérerons d'abord.

On dit qu'une variable aléatoire y est fonction (ou fonction certaine) d'une autre variable aléatoire x, s'il existe une application mesurable F de R dans R, telle que $y(\omega) = F(x(\omega))$ presque partout :

$$\int \{d\omega \mid \omega \in \Omega ; \; y(\omega) \neq F(x(\omega))\} = 0.$$

Bornons-nous au cas où x et y ont un moment d'ordre 2 : $x, y \in L_2$: alors x étant donné, l'ensemble H(x) des y qui sont fonction certaine de x est un sous-espace de Hilbert (sous-espace vectoriel fermé de L_2). Dans divers cas (notamment si x est bornée, si la loi de x est normale etc), H(x) est engendré par la suite des variables aléatoires $\{1, x, x^2, \ldots, x^n, \ldots\}$ (où on a noté x^n la fonction puissance n-ème : $x^n(\omega) = (x(\omega))^n$).

Soit $x, y, \in L_2$: x et y sont stochastiquement indépendantes si et seulement si les sous-espaces H(x) et H(y) se coupent orthogonalement suivant la droite des constantes. Avant de démontrer cela, rappelons les conditions géométriques d'orthogonalité : f est orthogonale dans L_2 à la droite des constantes si et seulement si $\langle f, 1 \rangle = 0$, i.e. si f est de moyenne nulle ; H(x) et H(y) se coupent orthogonalement suivant la droite des constantes si et seulement si est satisfaite l'une des trois conditions équivalentes suivantes :

1° $\forall u \in H(x), \forall v \in H(y) : \langle u,1 \rangle = \langle v,1 \rangle = 0 \Rightarrow \langle u,v \rangle = 0$

2° $\forall u \in H(x), \forall v \in H(y) : \mathrm{Cov}(u,v) = 0$

3° $\forall u \in H(x), \forall v \in H(y) : \langle u,v \rangle = \langle u,1 \rangle \, . \, \langle v,1 \rangle$.

(En effet, tous ces énoncés se ramènent à celui-ci : deux vecteurs orthogonaux à la droite constante et situés respectivement dans H(x) et H(y) sont orthogonaux entre eux). Ceci posé supposons x et y indépendants; soit u, v : $u = U(x)$, $v = V(y)$; on a :

$$\langle u,v\rangle = \int_\Omega u(\omega)\ v(\omega)\ d\omega = \int_{R^2} U(x)\ V(y)\ d(uv)^{-1}$$
$$= \int_{R^2} U(x)\ V(y)\ du^{-1}\ dv^{-1} = \int_\Omega u(\omega)\ d\omega \ \ x \int_\Omega v(\omega)\ d\omega :$$

la condition 3° est vérifiée . Supposons maintenant que x et y ne sont pas indépendants. Alors il existe deux intervalles réels A et B de R tels que :

$$\int \{d\omega \mid \omega \in \Omega;\ x(\omega) \in A\ ;\ y(\omega) \in B\} \neq$$
$$\int \{d\omega \mid \omega \in \Omega\ ;\ x(\omega) \in A\} \times \int \{d\omega \mid \omega \in \Omega\ ;\ y(\omega) \in B\}\ .$$

Désignons par U et V les fonctions caractéristiques de A et B respectivement (e.g. U vaut 1 sur A et est nulle sur R - A) ; soit u = U(x), v = V(y). L'inégalité ci-dessus peut être récrite comme suit :

$$\int_\Omega u(\omega)\ v(\omega)\ d\omega \neq \int_\Omega u(\omega)\ d\omega \times \int_\Omega v(\omega)\ d\omega\ ;$$

et la condition 3° n'est pas vérifiée. On a ainsi établi l'équivalence entre : x et y indépendants, d'une part, et H(x) et H(y) orthogonaux d'autre part.

Soit $\{x^1, \ldots, x^n\}$ une suite de n variables aléatoires : on a le sous-espace de Hilbert $H(\{x^1, \ldots, x^n\})$, ensemble des variables aléatoires u qui peuvent s'exprimer comme fonction composée $U(x^1(\omega), \ldots, x^n(\omega))$ des x^i.

Soit X une partie quelconque de L_2 : on a le sous-espace H(X), fermeture de la réunion des H(F) où F désigne une partie finie quelconque de X : $F = \{x^1, \ldots, x^n\} \subset X$.

Dans le cadre géométrique fourni par L_2, on va maintenant placer diverses notions de régressions. Soit $y \in L_2$, une variable aléatoire : on se propose d'approcher y en fonction d'autres variables aléatoires (généralement appelées : variables explicatrices). Soit $z \in L_2$, on considérera comme mesure de l'écart entre y et z le moment d'ordre 2 : $\|y - z\|^2 = \int_\Omega |y(\omega) - z(\omega)|^2\ d\omega$. C'est ce qu'on appelle faire l'approximation au sens des moindres carrés. Soit V une partie convexe fermée de L_2 : on sait qu'il existe dans V un v unique rendant minimum le carré de l'écart $\|y - v\|^2$; v est appelé la projection orthogonale de y sur V. Soit $\{x^1, \ldots, x^n\}$ un système fini de variables aléatoires : $V(\{x^1, \ldots, x^n\})$, ensemble des combinaisons linéaires des x^i, est un sous-espace fermé de L_2, dont la dimension est inférieure ou égale à n (inférieure à n si les x^i sont liées par une relation linéaire). Projeter orthogonalement y sur $V(\{x^1, \ldots, x^n\})$, c'est trouver la meilleure approximation de y en combinaison linéaire des x^i, ou encore faire une régression de y par rapport aux x^i : on a les coefficients a_i de la formule de régression $y \approx \Sigma a_i x^i$, en résolvant un système linéaire dont les coefficients sont les carrés de norme et les produits scalaires des vecteurs y et x^i (cf [Alg. Eucl.], TII B n° 12 § 2).

On aura la meilleure approximation de y en fonction quelconque des x^i en projetant orthogonalement y sur le sous-espace $H(\{x^1, \ldots, x^n\})$ ci-dessus défini. Nous parlerons ici de régression fonctionnelle. Si l'on accepte le modèle probabiliste, et qu'on suppose celui-ci exactement déterminé (hypothèses certes peu réalistes !, cf infra § 2) cette nouvelle approximation l'emporte en général de beaucoup sur la précédente : mais elle est définie par un système linéaire infini, en sorte que les applications pratiques de la régression fonctionnelle, comporteront nécessairement des développements limités.

Examinons particulièrement la régression de y en fonction d'une seule variable aléatoire x (à laquelle nous adjoindrons la constante 1). Notons :

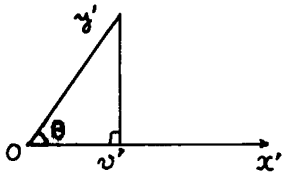
$$y = \text{Moy}(y) + y' \quad ; \quad x = \text{Moy}(x) + x'.$$

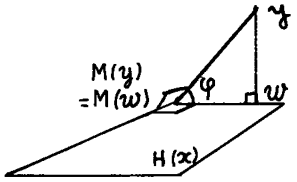
Pour la régression linéaire on a :

$$y \approx \text{Moy}(y) + \text{Var}(y)^{1/2}\,\text{Corr}(x,y)\,\text{Var}(x)^{-1/2}\,(x - \text{Moy}(x))$$

$$y \approx \text{Moy}(y) + v' = v$$

Notons $\text{Corr}(x,y) = \cos\theta$; la régression linéaire s'explique alors sur la figure ci-dessous :





où l'on voit de plus que :

$$\|y - v\|^2 = \|y' - v'\|^2 = \|y'\|^2 \sin^2\theta$$

$$= \text{Var}(y)\,(1 - \text{Corr}(x,y)^2)$$

La régression fonctionnelle s'effectue en projetant orthogonalement y sur H(x). Soit w la projection de y. Parce que H(x) contient la droite des constantes on a : Moy(y) = Moy(w) : cela résulte du théorème des trois perpendiculaires (la moyenne d'une fonction n'est autre que sa projection sur la droite des constantes).

Si l'on note φ l'angle aigu fait par le vecteur y' avec H(x) (i.e. l'angle des deux vecteurs $y' = y - M(y)$ et $w' = w - M(y)$) il vient :

$$\|y - w\|^2 = \|y' - w'\|^2 = \text{Var}(y)\,\sin^2\varphi$$

On a toujours : $\sin^2\varphi \leq \sin^2\theta$, car la régression fonctionnelle ne peut qu'améliorer la régression linéaire. On peut appeler $\cos\varphi$: coefficient de corrélation fonctionnelle de y avec x (ou, mieux, avec H(x)) : c'est le maximum de la corrélation entre y et une fonction (certaine) de la variable aléatoire x. On prendra garde que ce coefficient de corrélation ne dépend pas symétriquement de x et de y. Pour le voir, supposons que y est une variable aléatoire de moyenne nulle dont la loi est symétrique (e.g. une variable normale centrée) et que $x = y^2$. Toute fonction de x a un coefficient de corrélation (usuel) nul avec y : y' est orthogonal à H(x) ; tandis que x appartient évidemment à H(y). Pratiquement, on l'a dit, on ne peut effectuer de régression fonctionnelle, mais on effectuera, e.g., une régression polynomiale de degré n choisi : on projettera y sur le sous-espace vectoriel engendré par $\{1, x, \ldots, x^n\}$ (cf [Alg. Eucl.] § 2.2.2).

Pour conclure confrontons les notions de liaison, corrélation, régression linéaire, régression fonctionnelle. Si x et y ne sont pas liés, (sont stochastiquement indépendants) H(y) et H(x) sont orthogonaux, y et x sont non corrélés ($\text{Corr}(y,x) = 0$) et il n'y a pas de régression possible, même fonctionnelle, de y par rapport à x, sinon en posant $y \approx \text{Moy}(y)$. Si y et x ne sont pas stochastiquement indépendants, c'est qu'ils sont en un certain sens liés ; pourtant, on peut avoir $\text{Corr}(x,y) = 0$; et il se peut même (cf supra $x = y^2$) que la projection orthogonale de y sur H(x) soit Moy(y) : donc qu'il n'y ait pas de régression fonctionnelle possible (de y en fonction de x). Mais puisque H(x) et H(y) ne se coupent plus orthogonalement suivant la droite des constantes, on peut trouver dans H(x) et H(y), $u = U(x)$ et $v = V(y)$ ayant moyenne nulle et corrélés entre eux ($\text{Corr}(u,v) > 0$) : et c'est d'ailleurs une sorte de problème d'analyse factorielle que de déterminer de tels u, v de corrélation maxima (cf thèse P. Cazes analyse canonique). Enfin si Corr(y,x) est non-nul, la régression linéaire de y en fonction de x conduit à un résultat non trivial (à une approximation de y meilleure que Moy y). On sait que l'absence de corrélation : $\text{Corr}(y,x) = 0$, n'est qu'une condition nécessaire

d'indépendance, non une condition suffisante. Pourtant on assimile volontiers absence de corrélation à absence de liaison, sans songer que dans H(x) le sous-espace des éléments non corrélés à x est de codimension 1 (i.e. presque toute fonction de x est non corrélée à x !) : et y songe-t-on qu'il reste difficile d'exprimer avec justesse et concision, au terme d'une étude concrète les liens entre variables ou entre facteurs...

? *Les faits statistiques* : Quant aux applications, deux questions se posent. Le modèle probabiliste est-il conforme aux phénomènes étudiés? La connaissance que nous avons de ces phénomènes permet-elle de faire des calculs de probabilités ?

Nous l'avons dit ailleurs (cf Principes TII A n° 1 § 1°), il existe dans la nature trois types de lois probabilistes.

a) Les lois de symétrie.

b) Les lois ergodiques.

c) La mécanique quantique.

Dans les données soumises au statisticien, les symétries géométriques simples ne jouent pas un rôle important ; les phénomènes quantiques ne sont pas l'objet de cette étude. Il reste donc la possibilité que les processus complexes étudiés soient régis par des lois ergodiques, exprimant sous forme probabiliste l'équilibre de multiples causes. Le domaine des lois ergodiques est sans doute vaste : il s'en faut toutefois de beaucoup qu'on puisse concevoir l'existence de probabilités objectives partout où l'on fait usage de modèles probabilistes. De quoi nous donnerons deux raisons.

D'une part, on postule souvent qu'ont une probabilité d'être vraies des assertions mal définies. Exemple : probabilité qu'il fasse beau demain ;est-ce la probabilité qu'il fasse beau un quatorze juillet (j'écris le 13) en telle ville ? Ou la probabilité qu'au lendemain d'un jour d'été ensolleillé tel qu'aujourd'hui il n'éclate point d'orage ? A la limite, si l'on précise parfaitement les conditions d'aujourd'hui, demain aussi est rigoureusement fixé et il n'y a plus doute mais certitude.

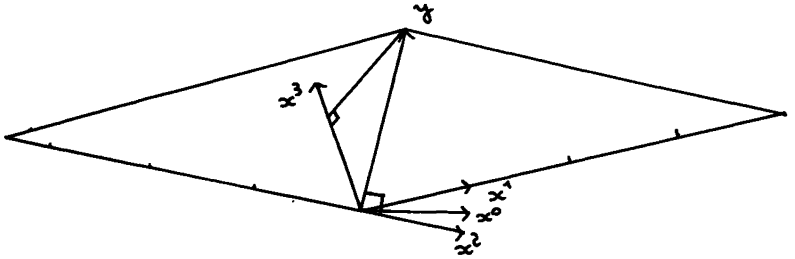
D'autre part, lors même qu'un domaine naturel est défini sans ambiguïté (e.g. probabilité pour qu'un artisan cordonnier laisse l'échope à son fils), les conditions peuvent se modifier trop vite pour que l'équilibre de causes infimes, présupposé par les lois ergodiques, soit jamais atteint. Peut-on faire un modèle théorique de l'économie française alors que, disons, le tiers des variations de prix est acquis au cours de crises qui, telles celle de 1968, ne relèvent pas d'un type général?

Mais les conceptions probabilistes suggèrent des opérations algébriques et permettent d'en évaluer la portée. C'est de ce point de vue que nous chercherons des bases pour la régression. Dans les §§ suivants nous proposons deux méthodes de calcul protégées contre les erreurs;ces méthodes (cf P. Cazes, thèse) ont déjà fait leurs preuves. Ici nous rappellerons seulement à quels résultats absurdes on peut être conduit si l'on calcule d'après des échantillons finis comme on le ferait légitimement sur des lois décrites mathématiquement.

Considérons les calculs de régression linéaire : ce sont des résolutions de systèmes linéaires dont les coefficients sont des covariances (ou des coefficients de corrélation) ou des moyennes. Covariances et moyennes ne sont connues, d'après un échantillon, qu'avec imprécision . Désignons par y une variable aléatoire pour laquelle on cherche une formule de régression linéaire ($y \approx w = \Sigma a_i x^i$) en fonction de n variables aléatoires $\{x^1, \ldots, x^i, \ldots, x^n\}$, (y compris la constante 1). En langage géométrique, dans l'espace L_2, $w = \Sigma a_i x^i$ est la projection orthogonale de y sur le sous-espace engendré par les x^i ; et les données (coefficients des équations) sont les produits scalaires (covariances) :

$$\{ < x^i , x^{i'} > | i, i' = 1, \ldots, n\} \quad ; \quad \{ < y, x^i > | i = 1, \ldots, n\} .$$

La projection orthogonale w est d'autant plus mal connue que le sous-espace $V(\{x^1,\ldots, x^n\})$ est moins bien déterminé par les x^i qui l'engendrent ; ce qui est le cas si p des x^i satisfont approximativement à une relation linéaire. L'exemple le plus simple est celui de la projection de y sur $V(\{x^1, x^2\})$, quand x^1 et x^2 sont presque colinéaires : du fait des erreurs, (analogues aux erreurs graphiques bien connues de ceux qui ont fait des épures), le plan V n'est guère mieux déterminé que comme un plan arbitraire contenant le vecteur x^1. Accidentellement, il se peut que y paraisse être contenu dans V, alors qu'en fait il lui est orthogonal ; on établira une formule de régression $y = a_1 x^1 + a_2 x^2$, dont les coefficients a_1 et a_2 seront très élevés, puisque x^1 et x^2 sont quasi-colinéaires, et qui n'aura aucun sens. Ci-dessous on a figuré un cas où $n = 3$: x^1 et x^2 ne diffèrent de x° que par des erreurs ; on a en fait : $\langle x^1,y\rangle=\langle x^2,y\rangle=0$; $\cos(x^3,y) = \sqrt{2/2}$. Il semble que y appartienne au plan $V(\{x^1,x^2\})$ et qu'on ait : $y = 4x^1 - 3{,}5x^2$; mais le plus sage serait d'estimer y par sa projection sur l'axe x^3 : $y \approx (2/3)x^3$, bien que $\cos(x^3,y) = \sqrt{2/2}$.



Il est instructif de présenter d'une autre manière le même argument. Expérimentalement, on ne connait pas l'espace probabilisé Ω, (cf § 1) ; on ne connait qu'un ensemble fini d'événements : soit $J \in \Omega$; card J=m. Les fonctions aléatoires y, x^i ne sont pas connues comme des fonctions sur Ω, mais seulement comme des fonctions sur J, dont les valeurs sont entachées d'erreurs : y est assimilé à $y^J = \{y(j) \mid j \in J\} \in R^J$. Les produits scalaires tels que $\langle y,x^i\rangle$ sont remplacés par des sommes finies, ou produits scalaires dans R^J (muni de la norme : moyenne des carrés des coordonnées) :

$$\langle y,x^i\rangle \approx \Sigma\{y(j)\, x^i(j)/m \mid j \in J\} = \langle y^J,x^{iJ}\rangle$$

C'est dans R^J qu'on calcule les projections orthogonales, qu'on cherche une expression linéaire approchée de y^J en combinaison linéaire des x^{iJ}. Considérons en particulier le cas extrême où $n \geq m$: si m est grand, (e.g. m = 200), les coefficients de corrélation et les moyennes, déterminés sur des échantillons nombreux, sont connus individuellement avec précision. Pourtant la formule de régression court grand risque d'être dépourvue de sens. En effet, hormis le cas exceptionnel où une relation linéaire exacte existe entre les n vecteurs x^{iJ}, ceux-ci engendrent R^J : y^J est donc une combinaison linéaire des x^{iJ}, même si la valeur exacte de tous les produits scalaires $\langle y,x^i\rangle$ est zéro. On a toujours une formule de régression qui paraît excellente, mais celle-ci peut n'être calculée que sur les fluctuations de quantités nulles. Ici la cause fondamentale d'erreur est qu'on tente de faire une régression par rapport à autant ou plus de variables x^i qu'il n'y a d'individus (ou d'événements) j ;

on se protégerait en réduisant le nombre n des variables explicatrices par une analyse factorielle.

Le cas le plus absurde est celui de la régression polynomiale poussée au dernier degré. Soit y et x deux variables aléatoires indépendantes. Supposons que sur un ensemble J d'événements la variable x prenne des valeurs x(j) toutes deux à deux distinctes. Alors toute fonction z sur l'ensemble J peut être considérée comme une fonction composée de la fonction x ; et ce d'une infinité de manières, car on a : $z(j) = Z(x(j))$, où Z désigne une fonction réelle d'une variable réelle dont les valeurs z(j) ne sont déterminées que pour les valeurs x(j) de la variable. En particulier il existe un polynôme d'interpolation de degré m - 1, (m = card J), tel que :

$$\forall\, j \in J : y(j) = \Sigma\{a_i\, x(j)^i \mid i = 0, 1, \ldots, m-1\}.$$

Formule de régression polynômiale qui est absurde car, on l'a dit, y et x sont indépendants en probabilité.

Dans la suite, nous nous efforcerons de protéger les calculs de régression contre les résultats absurdes dont on a donné ici les exemples.

3 Régression linéaire par rapport à des variables entachées d'erreurs:

On a vu au § 2 qu'une combinaison linéaire de variables explicatives affectées de forts coefficients et se soustrayant entre elles, peut n'être qu'un conglomérat d'erreurs dont la coïncidence avec la variable à expliquer est toute fortuite. D'où l'idée de protéger la régression en bornant les valeurs absolues des coefficients. Au § 4 on considère, d'après P. Cazes, le cas où, les variables aléatoires étant essentiellement positives, il est naturel aussi d'imposer que les coefficients de régression soient tous positifs ; ce qui élimine radicalement les combinaisons soustractives dépourvues de sens. Ici nous nous plaçons dans le cadre d'un modèle d'erreurs assez commun, qui sans mériter en lui-même une confiance absolue, suggère une règle pour borner les valeurs absolues des coefficients de régression.

Soit I un ensemble fini de variables aléatoires. Posons que la valeur x^i de la i-ème variable est la somme des deux variables aléatoires: un terme qu'on appellera vrai et notera v^i et une erreur de moyenne nulle, ou reste, noté r^i. Admettons que les r^i soient indépendants entre eux et indépendants des v^i ; et désignons par $(\varepsilon^i)^2$ la variance de r^i, variance que dans la suite nous supposerons avoir estimée, (par exemple en comparant des mesures répétées sur un même événement). Si, en particulier, l'une des variables considérées, x°, est la constante 1, on a : $x^\circ = v^\circ = 1$, $r^\circ = 0 = (\varepsilon^\circ)^2$.

Chercher une formule de régression linéaire de l'une des variables v^i en fonction des mesures des autres :

$$v^i \approx \Sigma\{a_{i'}\, x^{i'} \mid i' \in I\ ;\ i' \neq i\},$$

revient à chercher une combinaison linéaire $z = \Sigma\{a_{i'}\, x^{i'} \mid i' \in I\}$ dont le moment d'ordre 2 soit minimum, sous la contrainte $a_i = -1$: le moment d'ordre 2 de z sera le moment d'ordre 2 de l'erreur sur v^i (erreur commise en assimilant v^i à $\Sigma\{a_{i'}\, x^{i'} \mid i' \neq i\}$), augmenté de $(\varepsilon^i)^2$. Nous sommes ainsi conduits au problème suivant : d'après un ensemble fini J d'observations (card J = m) (i.e. un tableau de nombres $x^{JI} = \{x^{ji} \mid j \in J, i \in I\}$, où x^{ji} est la valeur mesurée de la variable i sur l'événement j) estimer le moment d'ordre 2 d'une combinaison linéaire $z = \Sigma a_i\, x^i$ (les variances d'erreurs $(\varepsilon^i)^2$, étant, répétons-le, fixées une fois pour toutes).

D'une part, on peut estimer $Mt^2(z)$ sur l'ensemble J d'observations:

$$Mt^2(z) \approx (1/m)\ \Sigma_j\ (z^j)^2 =$$
$$\frac{1}{m}\Sigma\{(\Sigma\{a_i x^{ji} | i \in I\})^2 \mid j \in J\} = q(a_I, a_I)\ ;$$

où on a noté q la forme quadratique (forme bilinéaire symétrique) sur R_I dépendant du tableau x^{JI} des observations, et dont les coefficients sont :

$$q^{i'i''} = (1/m)\ \Sigma\{x^{ji'}\ x^{ji''} \mid j \in J\}\ .$$

D'autre part, la variance de z est bornée inférieurement par la variance d'erreur :

$$Var(z) \geq \Sigma\{(\varepsilon^i)^2\ a_i^2 \mid i \in I\} = \varepsilon^2(a_I, a_I)\ ;$$

d'où pour $Mt^2(z)$ la majoration :

$$Mt^2(z) \geq \varepsilon^2(a_I, a_I) + Moy(z)^2,$$

où l'on peut estimer Moy(z) par $(1/m)\ \Sigma z^j$. Ce que suggère l'estimation suivante :

$$Mt^2(z) \approx \sup(q(a_I,a_I),\ \varepsilon^2(a_I, a_I) + (\Sigma_j\ z^j/m)^2).$$

Sous des hypothèses de normalité, (hypothèses rarement justifiables, mais commodes pour justifier des calculs), on trouve cette estimation au terme d'un calcul de maximum de vraisemblance, que nous donnons ici à titre d'exercice. Cherchons parmi toutes les lois normales de moyenne quelconque b, et de variance σ^2 supérieure ou égale à $\varepsilon^2(a_I, a_I)$ celle qui a le plus de chances de conduire à l'échantillon des valeurs observées pour z. On doit rendre maximum le produit des densités.

$$\Pi\{(2\pi)^{-1/2}\ \sigma^{-1}\ \exp(-(z^j - b)^2/(2\sigma^2)) \mid j \in J\}\ ;$$

ou encore :

$$\sigma^{-m}\ \exp((-\Sigma_j(z^j)^2 + 2\,b\,\Sigma_j\ z^j - m\,b^2)/2\sigma^2).$$

Pour σ donné, la valeur optima de b est celle qui rend maximum :

$$2\,b\,\Sigma_j\ z^j - m\,b^2\ ;\ \text{d'où :}\ b = \Sigma_j\ z^j/m \approx Moy\ z.$$

La valeur optima de σ^2 est maintenant celle qui, sous la contrainte d'être supérieure à ε^2, rend maximum :

$$\sigma^{-m}\ \exp((-\Sigma_j\ (z^j)^2 + m\,b^2)\ /\ 2\sigma^2)$$

en dérivant logarithmiquement il vient :

$$\sigma^2 = \sup(\varepsilon^2,\ (\Sigma_j(z^j)^2/m) - b^2)\ ;$$

d'où pour $Mt^2(z)$, équivalent à $b^2 + \sigma^2$, l'estimation suggérée plus haut. (On pourrait objecter que la variance d'une variable normale est systématiquement sous-estimée par les calculs de maximum de vraisemblance ; mais d'une part l'erreur relative étant de 1/m, peut être négligée ici ; d'autre part nous nous intéressons non à la variance, mais au moment d'ordre 2, pour lequel en tout état de cause $(1/m)\Sigma_j(z^j)^2$ est le meilleur estimateur sans biais).

Revenons maintenant au problème de régression : sous la contrainte $a_i = -1$ (pour un i donné), trouver a_I rendant minimum l'estimateur adopté ci-dessus pour $Mt^2(\Sigma a_i\ x^i)$. Remarquons d'abord que si la constante 1 se trouve parmi les x^i, (soit $1 = x^\circ$), l'optimum correspond certainement à une valeur de $a_\circ$ telle que $\Sigma_j\ z^j = 0$: car les autres a_i restant constants, ce choix de $a_\circ$ rend minimum à la fois $q(a_J, a_J)$ et $\varepsilon^2 + (\Sigma\ z^j/m)^2$. Dans la suite, simplifiant ainsi l'écriture sans restreindre le domaine des applications, nous supposerons que les variables mesurées x (tant la variable à expliquer x^i, que les variables explicatrices $\{x^{i'} | i' \neq i\}$) ont été réduites, par addition d'une constante, à avoir moyenne nulle et que

la constante 1 n'est plus parmi les x. L'on doit donc trouver, sur l'hyperplan $H_i = \{a_I | a_I \in R_I \ ; \ a_i = -1\}$, le minimum du sup des deux formes quadratiques que nous avons notées $q(a_I , a_I)$ et $\varepsilon^2(a_I , a_I)$. Avant de chercher un algorithme d'approximation, considérons le problème géométrique. La forme quadratique ε^2 est définie positive (car les $(\varepsilon^i)^2$ sont tous positifs, toute mesure étant entachée de quelque erreur) ; dans H_i , $\varepsilon^2(a_I , a_I)$ est minimum au point $-\delta^i_I$ ($a_i = -\delta^i_i = -1$; et les autres $a_{i'} = -\delta^i_{i'} = 0$).

Si on munit H_i de la structure euclidienne définie par ε^2, on peut se représenter les surfaces (plus exactement les ensembles) de niveaux de $\varepsilon^2(a_I , a_I)$ dans H_i , comme des sphères (ou hypersphères) concentriques, de centre $(-\delta^i_I)$. La forme quadratique q est positive ; et elle est généralement définie positive, sauf au cas exceptionnel où certains des x^i satisferaient exactement une relation linéaire. Dans H_i la forme $q(a_I,a_I)$ atteint son minimum en un point α_I qui n'est autre que la solution du problème de régression linéaire usuel (au cas exceptionnel où q est dégénérée, le minimum dans H_i est atteint sur toute une sous-variété linéaire, que nous noterons A_i : $A_i \subset H_i$). Les ensembles de niveaux de q sont dans H_i des ellipsoïdes concentriques de centre α_I. (Si q est dégénérée, on a des cylindres elliptiques dont les génératrices sont parallèles à A_i, et dont les sections transverses ont leur centre sur A_i , axe du cylindre ; désormais nous nous restreindrons à la sous-variété linéaire H'_i issue de $(-\delta^i_I)$ dans H_i perpendiculairement à A_i en un point que nous noterons α_I : le minimum cherché se trouve certainement dans H'_i , où le problème se retrouve tel qu'il est dans H_i si q n'est pas dégénérée).

Si $\varepsilon^2(\alpha_I , \alpha_I) < q(\alpha_I , \alpha_I)$, le minimum cherché est atteint en α_I : la régression linéaire usuelle est acceptable. Le cas :

$$\text{Var } x^i = q(-\delta^i_I , -\delta^i_I) < \varepsilon^2(-\delta^i_I , -\delta^i_I) = (\varepsilon^i)^2,$$

n'est pas à envisager, car la variance de l'erreur sur la variable x^i ne peut être postulée inférieure à la variance empirique de x^i, évaluée sur l'échantillon J. La méthode de régression proposée ici ne différera donc de la régression linéaire usuelle que dans le cas où l'on a simultanément :

$$q(\alpha_I , \alpha_I) < \varepsilon^2(\alpha_I , \alpha_I) \ ;$$
$$(\varepsilon^i)^2 = \varepsilon^2(-\delta^i_I , -\delta^i_I) < q(-\delta^i_I , \delta^i_I)$$

le minimum cherché est alors supérieur à la fois à $(\varepsilon^i)^2$ et à $q(\alpha_I , \alpha_I)$; et il est atteint en un point β_I situé en quelque sorte entre α_I et $-\delta^i_I$. Cherchons β_I.

Pour toute quantité positive η^2 notons :

$$B(\eta^2) = \{a_I | a_I \in H_i \ ; \ \varepsilon^2(a_I , a_I) \leq \eta^2\}$$
$$E(\eta^2) = \{a_I | a_I \in H_i \ ; \ q(a_I , a_I) \leq \eta^2\}$$

Si $\eta^2 < (\varepsilon^i)^2$, $B(\eta^2)$ est vide ; sinon c'est une boule de centre $(-\delta^i_I)$ et de rayon $(\eta^2 - (-\varepsilon^i)^2)^{1/2}$. Si $\eta^2 < q(\alpha_I , \alpha_I)$, $E(\eta^2)$ est vide, sinon c'est un ellipsoïde (solide, plein) de centre α_I. (Les termes distincts de sphère

et d'ellipsoïde ne doivent pas nous masquer le fait que les deux formes q et ε^2 jouent dans notre problème des rôles symétriques. Si nous avons choisi plutôt ε^2 pour métrique euclidienne c'est d'une part parce qu'elle s'écrit simplement comme une somme de carrés, d'autre part parce qu'elle n'est pas comme q sujette à dégénérescence. Quant à l'usage même du langage euclidien, il nous aide à bénéficier ici de notre expérience de l'espace).

Soit m^2 défini par :

$$m^2 = \inf\{\eta^2 \mid \eta \in R \;\; ; \;\; B(\eta^2) \cap E(\eta^2) \neq \emptyset\} \; ;$$

m^2 est le minimum, dans H_i, de $\sup(q, \varepsilon^2)$; et ce minimum est atteint en un point unique β_I : $\{\beta_I\} = B(m^2) \cap E(m^2)$. Cette construction peut s'expliquer ainsi : si $B(\eta^2) \cap E(\eta^2) = \emptyset$, on a partout dans H_i $\sup(q, \varepsilon^2) > \eta^2$; on fait donc croître η^2 jusqu'à ce que $B(\eta^2)$ et $E(\eta^2)$ se touchent ; et ce premier contact a lieu en un point unique car boule et ellipsoïde sont des convexes sans facette linéaire. De plus, β_I se trouve sur la quadrique S (hypersurface de H_i) d'équation :

$$S = \{a_I \mid a_I \in H_i \;\; ; \;\; q(a_I, a_I) = \varepsilon^2(a_I, a_I)\} \; ;$$

quadrique qui possédera généralement des nappes infinies ; et β_I est le point de S le plus proche de $(-\delta^i_I)$ (car, sur S, $\sup(q, \varepsilon^2) = \varepsilon^2$ n'est autre que le carré de la distance à $(-\delta^i_I)$, augmenté de la constante $(\varepsilon^i)^2$). Parmi les pieds des perpendiculaires issues du point $(-\delta^i_I)$ à la quadrique S, β_I est le seul qui soit intérieur à la sphère de diamètre $[\alpha_I, -\delta^i_I]$. En effet soit $\gamma_I \in S$ tel que $\gamma_I + \delta^i_I$ soit le vecteur normal à S en γ_I ; en γ_I sont tangents deux convexes $E(\eta^2)$ et $B(\eta^2)$; ces deux convexes sont situés de part et d'autre de leur plan tangent commun en γ_I si et seulement si l'angle formé en γ_I par les deux segments $[\gamma_I, \alpha_I]$ et $[\gamma_I, -\delta^i_I]$ est obtu ; c'est-à-dire si γ_I appartient à la sphère de diamètre $[\alpha_I, -\delta^i_I]$, et alors γ_I n'est autre que β_I. Cette propriété aidera à suivre l'algorithme de recherche de β_I.

Signalons ici une particularité géométrique, dont il est toutefois difficile de faire usage pratiquement : β_I se trouve sur la courbe C lieu des pieds des perpendiculaires issues de $(-\delta^i_I)$ aux ellipsoïdes $E(\eta^2)$. Il est classique que cette courbe admet une équation paramétrique simple, que nous allons rappeler ici. Supposons que l'on ait pour $a_I \in H_i$:

$$q(a_I, a_I) = q(\alpha_I, \alpha_I) + \Sigma\{(a_{i'} - \alpha_{i'})^2 / b^2_{i'} \mid i' \in I'\} \; ;$$

où on a noté $I' = I - \{i\}$: on peut toujours se ramener à ce cas en prenant pour axes de coordonnées les axes de la forme q (dans la métrique ε^2). La normale en un point a_I à la surface de niveau de q (dans H_i) a pour vecteur directeur :

$$n_{I'} = \{(a_{i'} - \alpha_{i'})/b^2_{i'} \mid i' \in I'\}$$

cette normale passe par $(-\delta^i_I)$ si les vecteurs n et $(a_I + \delta^i_I)$ sont proportionnels. D'où l'équation de la courbe C en fonction du paramètre réel t :

$$\forall\, i' \in I' : (a_{i'} - \alpha_{i'})/b^2_{i'} = t\, a_{i'} \; ;$$
$$\forall\, i' \in I' : \; a_{i'} = \alpha_{i'}/(1 - t\, b^2_{i'}) .$$

Cette équation paramétrique est très simple et semble devoir permettre de définir β_I comme le point d'intersection de $C \cap S$ le plus proche

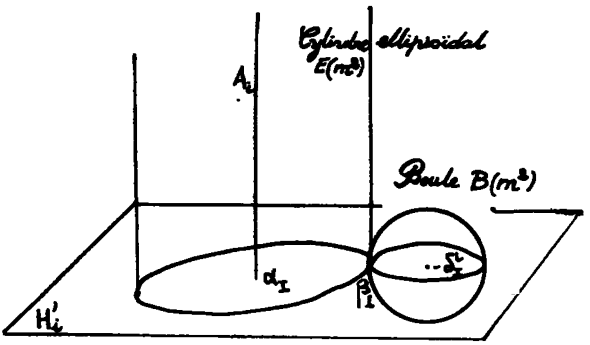


Figure 3.1 : le cas singulier où q est dégénérée

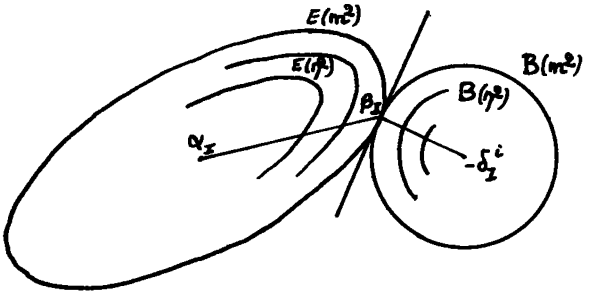


Figure 3.2 : le point β_I est un contact de $B(m^2)$ et $E(m^2)$; on notera que l'angle $\alpha\beta\delta$ est droit.

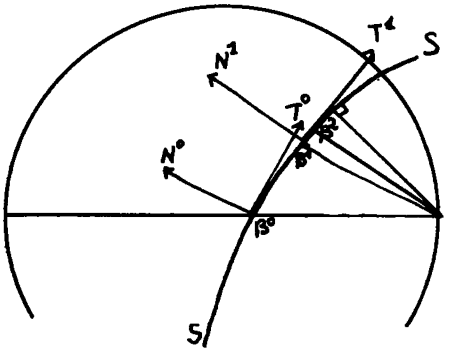


Figure 3.3 : l'algorithme itératif de construction de β_I

de $-\delta^i_I$. Mais dans la pratique l'on n'est pas dans un système d'axes favorable ; un changement d'axes semble coûteux ; dans des axes quelconques, l'équation de la courbe C est :

$$\forall\, i' \in I' : \Sigma\{q^{i'i''}(a_{i''} - \alpha_{i''}) \mid i'' \in I'\} = t\, a_{i'} ;$$

et il n'est pas facile d'avoir en fonction explicite de t la solution de ce système linéaire.

Nous sommes maintenant à même de suggérer un algorithme itératif de calcul pour β_I. Ici, nous identifierons H_i à $R_{I'}$: en effet tout point a_I de H_i a(-1) pour i-ème composante, on peut donc se borner à $a_{I'} = \{a_i \mid i \in I'\}$; ainsi le point $-\delta^i_I$ est identifié à l'origine $O_{I'}$ de $R_{I'}$ (vecteur dont toutes les coordonnées sont nulles). Dans $R_{I'}$ un point $a_{I'}$ de la quadrique S satisfait à l'équation :

$$\Sigma\{q^{kh}\, a_k\, a_h \mid k,h \in I'\} - 2\Sigma\{q^{ih} a_h \mid h \in I'\} + q^{ii} - (\varepsilon^i)^2 - \Sigma\{(\varepsilon^h)^2 a_h^2 \mid h \in I'\} = 0$$

c'est l'équation $q(a_I, a_I) - \varepsilon^2(a_I, a_I) = 0$, où l'on a remplacé a_i par sa valeur -1 . Nous noterons désormais $S(a_{I'})$ le polynôme inhomogène du second degré, équation de S. Dans $R_{I'}$, muni de la structure euclidienne définie par ε^2, le gradient fournit un vecteur normal à S au point $a_{I'}$:

$$N_{I'}(a_{I'}) = \{N_{i'} \mid i' \in I'\}; N_{i'} = (\varepsilon^{i'})^{-2}\, \partial S(a_{I'})/\partial a_{i'} .$$

On cherche, à l'intérieur de la sphère de diamètre $O\alpha$, (i.e. $[-\delta^i_I, \alpha_I]$), le point β de S le plus proche de O. Pour cela on construit une suite $\beta^0, \beta^1, \ldots \beta^n$. Le point β^0 est l'intersection avec S du segment $O\alpha$ (cette intersection existe et est unique, sous les hypothèses faites quant aux valeurs de q et ε^2 en $O = -\delta^i_I$ et α). Reste à dire comment on passe de β^n à β^{n+1}. On se place dans le plan $P^n = [O\, \beta^n, N^n]$ défini par le segment $O\beta^n$ et la normale en β^n à S. Il serait agréable de prendre pour β^{n+1} le point de la conique $P^n \cap S$, (courbe ordinaire dans le plan $[O\beta^n, N^n]$) le plus proche de O ; mais le calcul précis d'un tel point est coûteux. Le plus simple est sans doute de se déplacer sur S à partir de β^n dans la direction T^n intersection de P^n et de l'hyperplan tangent à S :

$$T^n = -\beta^n + (<\beta^n . N^n> / \|N^n\|)\, N^n$$

(où on a noté β^n le vecteur $O\beta^n$ et où produit scalaire et norme s'entendent au sens de la métrique ε^2) ; on prendra un point $\beta^{n+1} = (1-u)\beta^n + tT^n$; la quantité du premier ordre t, et la quantité du second ordre u étant choisies de sorte qu'en restant sur S on se rapproche de O.

4 Régression linéaire à coefficients positifs : Comme bien d'autres commodités mathématiques, l'usage des nombres algébriques peut faire oublier la nature des réalités qu'il permet de saisir. En elle-même toute quantité n'est-elle pas positive! Ce que A apporte à B n'est-il pas positif et proportionnel à A? Amoins que A ne se nourrisse aux dépens de B ? On conçoit donc qu'il soit souvent naturel d'imposer d'être positifs, aux coefficients d'une régression par rapport à des quantités positives: du même coup les coefficients de régression se trouveront bornés et, on l'a dit, la régression sera protégée. L'algorithme d'approximation au sens des moindres carrés sous contrainte positive, étant exactement décrit par P. Cazes qui l'a mis au point, nous nous bornerons ici à des principes géométriques et à des exemples.

Considérons la variable à estimer y et les variables explicatrices x^i, (y compris éventuellement la constante 1) comme des fonctions réelles

sur un ensemble fini J d'observations, c'est-à-dire comme des vecteurs de R^J : y^J , x^{Ji} ; et supposons R^J muni d'une métrique euclidienne dite métrique d'erreur (qui sera généralement la racine carrée de la somme des carrés des coordonnées ; à moins qu'on ne doive pondérer inégalement les divers j, ce qui sera le cas, (cf infra), si les x^{ji} sont des fréquences). En général, Card I < Card J : il y a moins de variables que d'observations ; les vecteurs $\{x^{Ji} | i \in I\}$ forment un système linéairement indépendant dans R^J; et l'ensemble des combinaisons linéaires positives, noté $V^+(x^{JI})$, est un orthant convexe (e.g. si Card I = 3 on a un trièdre) du sous-espace vectoriel $V(x^{JI})$, engendré par les vecteurs $\{x^{Ji} | i \in I\}$ dans R^J :

$$V^+(x^{JI}) = \{\Sigma\{a_i \; x^{Ji} \mid i \in I\} | \{a_i \mid i \in I\} \in R_I^+\} .$$

La recherche du point V^+ le plus proche de y^J pourra se faire en deux étapes : on projettera y^J sur $V(x^{JI})$ orthogonalement (au sens de la métrique de R^J), suivant un point w^J qui n'est autre que le résultat d'une régression linéaire usuelle, sans contrainte. Puis dans l'espace vectoriel $V(x^{JI})$, de dimension Card I , muni de la métrique induite par celle de R^J, on cherchera le point w^{+J} de $V^+(x^{JI})$ qui soit le plus proche de w^J. Ce problème résolu par P. Cazes rentre dans le cas usuel de la recherche d'un optimum sur un convexe ; la difficulté étant d'avoir un algorithme de rapidité suffisante.

Une première classe d'application sur laquelle il n'y a pas lieu d'insister est celle d'une régression entre variables essentiellement positives, où l'on présume que les variables explicatives x^i ne détruisent pas la quantité à expliquer y , mais au contraire y contribuent positivement. On pose donc : $y \approx \Sigma a_i \; x^i$; $a_i \in R^+$. C'est ainsi que P. Cazes explique le taux y d'un oligo-élément (e.g. le manganèse) dans divers échantillons de roche, par les apports de divers minéraux majeurs (e.g. le calcaire, l'argile etc).

Plus spécifique est l'application aux processus physiques régis par une équation intégrale à noyau positif, cadre dans lequel se placent plusieurs cas particuliers intéressants. Supposons que l'on ait la formule théorique (ou modèle) :

$$f(x) = \int N(x,y) \; g(y) \, dy ,$$

où N, f, g sont essentiellement positifs ; et que , la fonction f(x) étant connue expérimentalement (soit par un nombre fini de ses valeurs, soit par un tracé automatique), on désire estimer g(y). En subdivisant l'intervalle d'intégration et éliminant éventuellement les extrémités infinies qui ne contiennent que des masses négligeables (Il n'est pas nécessaire de prendre les mêmes subdivisions pour les variables x et y : on intégrera en faisant usage d'une suite d'intervalles consécutifs égaux en longueur ; puis on comparera le résultat à f(x) d'après une subdivision en intervalles de même masse) on obtient un système fini (I et J sont des ensembles finis) :

$$\forall \; j \in J \; : \; f_j = \Sigma_{i \in I} \; N_j^i \; g_i ,$$

où les N_j^i sont fournis par le modèle théorique, les f_j sont des données expérimentales, et les g_i sont des inconnues à estimer, tous ces nombres étant essentiellement positifs. Si l'on écrit vectoriellement :

$$f_J = \Sigma\{g_i \; N_J^i \mid i \in I\} \in R_J$$

on voit qu'il ne s'agit de rien d'autre que de trouver la meilleure approximation d'un vecteur f_J en combinaison linéaire positive des vecteurs N_J^i.

Quant à la métrique d'erreur dont est muni R_J, le choix en est plus ou moins sûr. Si les f_j sont des fréquences, ou résultent de comptages (f_j est nombre de fois qu'un certain phénomène s'est manifesté sur le i-ème intervalle de l'axe des x) il s'impose de prendre la métrique du χ^2 :

$$\|X_J\|^2 = \Sigma\{(X_j)^2/f_j \mid j \in J\} ,$$

en se gardant, comme toujours dans les calculs de χ^2, des classes d'effectifs trop faibles.

Un exemple relatif à des courbes de fluorescence, traité par Messieurs Pagès et Turpin au C.E.A., est exposé dans la thèse de P. Cazes. Une autre application est le problème inverse du lissage (le cas particulier d'une courbe chromatographique a été rencontré par Talfer à la Cie de Pont à Mousson). La fonction mesurée f(x) est reliée à une fonction sous-jacente g par une diffusion de noyau N. Pour y donné, N(x,y) considéré comme fonction de x est ici une courbe en cloche, d'aire totale 1, centrée approximativement en y. Si en particulier N(x,y) = N(x-y,O)= K(x-y), on a affaire à une équation de convolution : f = K * g. Ce qui suggère une dernière application à un problème d'ajustement : la décomposition d'une densité de probabilité empirique en une somme finie d'un nombre aussi petit que possible de densités normales. (Sur ce problème,voir l'article [Décomposition] où sont exposées les recherches de P. Cazes et A. Schroeder).

Notons, comme il est d'usage pour une fonction de Dirac $\delta(x-a)dx$, la masse unité placée au point d'abscisse a de la droite réelle. Notons:

$$N(x - a ;\sigma^2) = \sigma^{-1}(2\pi)^{-1/2} \exp(-(x - a)^2/(2\sigma^2))$$

la densité de la loi normale de moyenne a et variance σ^2. On sait que l'on a les formules de convolution :

$$\delta(x - a) * \delta(x - b) = \delta(x - a - b)$$

(masse en a convolée avec masse en b = masse en a+b : c'est la définition de la convolution):

$$\delta(x - a) * N(x - b ; \sigma^2) = N(x - a - b ; \sigma^2)$$

$$N(x - a ; \sigma^2) * N(x - b ; \tau^2) = N(x - a - b ; \sigma^2 + \tau^2)$$

(la convolée de deux lois n'est autre que la loi de la somme de deux variables aléatoires indépendantes dont chacune suit respectivement l'une des lois données : nous ne faisons donc que rappeler, que la somme de deux variables laplaciennes indépendantes, a pour moyenne la somme des moyennes et variance la somme des variances. On notera que l'on peut poser $\delta(x) = N(x;O)$ et se restreindre à la dernière formule). Le problème de décomposition peut s'énoncer comme suit : la loi empirique f(x) dx étant donnée, trouver une formule d'approximation de la forme :

$$f(x) \approx \Sigma_{j\in J} b_j . N(x - a_j ; \sigma_j^2) = f'(x)$$

comportant aussi peu de termes que possible. L'énoncé classique est vague, nous proposons la stratégie précise suivante. Supposons que dans la formule ci-dessus les σ_j^2 aillent en croissant $\sigma_1^2 < \sigma_2^2 \ldots$, f'(x) étant décomposée en somme d'une suite de pics de plus en plus larges. Soit $\sigma^2 \le \sigma_1^2$: on a :

$$f'(x) = N(x;\sigma^2)\ \Sigma\{(b_j\ N(x - a_j ; \sigma_j^2 - \sigma^2) \mid j \in J\} ;$$

il est possible de faire la déconvolution par $N(x;\sigma^2)$ tant que $\sigma^2 \le \sigma_1^2$. Mais pour $\sigma^2 = \sigma_1^2$ le premier terme du quotient de convolution (la somme à droite) est $b_1\ N(x - a_1 ; O) = b_1\ (x - a_1)$; c'est une masse b_1 placée en a_1. On essaiera donc par la méthode des moindres carrés sous contrainte positive, de diviser la loi empirique donnée f(x) par des $N(x;\sigma^2)$ pour σ^2 de plus en plus grand. On doit s'arrêter quand, selon le critère du χ^2 ,

l'écart entre f(x) et l'expression approchée de la forme $N(x;\sigma^2) * g$ atteint la limite du vraisemblable. La densité g présente alors un pic (ou exceptionnellement plusieurs) correspondant au $b_1\ \delta(x - a_1)$ du cas-modèle. On retranche ce pic de g, soit g_1 le reste : on opère sur g_1 comme on l'a fait sur f ; et ainsi de suite jusqu'à ce que selon le critère du χ^2 le reste soit de poids négligeable. On a alors obtenu :

$$f \approx N(x;\sigma_1^2) * (b_1\ \delta(x - a_1) + g_1)$$

$$g_1 \approx N(x;\tau_2^2) * (b_2\ \delta(x - a_2) + g_2)$$

$$g_2 \approx N(x;\tau_3^2) * (b_3\ \delta(x - a_3) + g_3) \ldots$$

d'où :

$$f \approx b_1 N(x-a_1\ ;\ \sigma_1^2) + b_2\, N(x-a_2;\sigma_1^2+\tau_2^2) + b_3\, N(x-a_3\ ;\ \sigma_1^2 + \tau_2^2 + \tau_3^2) + \ldots$$

C'est, croyons-nous, la formule de décomposition cherchée.

5 *Remarque* : La méthode de l'échantillon d'épreuve utilisée par J.-M. Romeder en analyse discriminante (cf [Sep. Corr.] § 6) pourrait servir en régression : une formule de régression peut, comme une formule de discrimination qui en est un cas particulier (le cas où la variable à expliquer prend ses valeurs dans un ensemble fini ; par exemple un ensemble à deux éléments s'il s'agit de discriminer entre deux classes) être établie d'après une partie seulement des individus disponibles ; le reste de ceux-ci (dit : échantillon d'épreuve) étant réservé pour éprouver la validité de la formule.