BASES GÉOMÉTRIQUES DE LA THÉORIE DES CORRÉLATIONS

PAR

R. FERON

Correspondant du Centre de Statistique mathématique

Résumé

Etant donné un couple aléatoire XY de fonction de répartition $F(x, y)$ Serge BERNSTEIN a montré que si Y est en corrélation dure par rapport à x et X en corrélation dure par rapport à y et si $F(x, y)$ est de classe C^5, alors ou bien X et Y sont indépendantes ou bien le couple XY obéit à une loi de Bravais. Nous montrons que, si les lignes de régression sont linéaires, la conclusion est valable quelle que soit $F(x, y)$. D'autre part, nous nous servons de la notion de la corrélation dure de S. BERNSTEIN et de la notion de corrélation isogène de SARMANOY pour associer à tout couple aléatoire XY un couple plus simple X*Y* sur lequel les notions de régression et de corrélation ont une interprétation géométrique très simple. On retrouve ainsi notamment r^2, η^2 et les indices de connexion de Gini.

Summary

Let X Y be a given pair of random variables and F(x, y) their distribution function. S.N. Bernstein has shown that if Y is in hard correlation with x and if X is in hard correlation with y and if $F(x, y)$ belongs to the class C^5, then, either X and Y are independant or the pair XY follows a Bravais'law. We show here that if the régression curves are linear, then, the conclusion is true whatever $F(x, y)$ may be. On the other hand, we use Berstein's notion of hard correlation and Sarmanov's notion of isogenous correlation to associate with every pair of random variables XY another simplier pair X* Y* on which the notions of regression and correlation have a very simple geometric interpretation. Thus, we find again, specially, r^2, η^2 and Gini's connexion index.

Considérons un couple aléatoire XY de fonctions de répartition $F(x, y)$.

Etendant légèrement les définitions de S. BERNSTEIN qui ne considère que les cas des couples aléatoires possèdant une densité de probabilité, nous dirons que Y est en corrélation dure par rapport à x si il existe un système admissible de fonctions de répartition liées $F_x(y)$ tel que, quel que soit x, on ait $F_x(y) = C[y - \alpha(x)]$ ou C est une certaine fonction de répartition fixe.

S. BERNSTEIN montre alors que si les dérivées des 5 premiers ordres de $F(x, y)$ existent et si Y est en correlation dure par rapport à x et X en corrélation dure par rapport à y, alors, ou bien X et Y sont indépendantes ou bien le couple XY obéit a une loi normale.

J'ai montré que si on suppose les lignes de régression linéaires on peut se libérer de toute restriction conservant la dérivabilité. Soit en effet (*cf.* FERON et FOURGEAUD 1) $\phi(u, v)$ la fonction caractéristique du couple et $\phi_X(v)$ la fonction caractéristique associée à $F_x(y)$ on a :

$$\phi(u, v) = \int \phi_x(v) e^{iux} dF(x, \infty) \quad (1)$$

$$\phi(o, v) = \int \phi_x(v)\, dF(x) \quad (2)$$

Soient $y = ax$ et $x = by$ les lignes de régression de Y en x et de X en y on aura :

$$\phi_x(v) = e^{iax\, v} \phi_1(v) \quad (3)$$

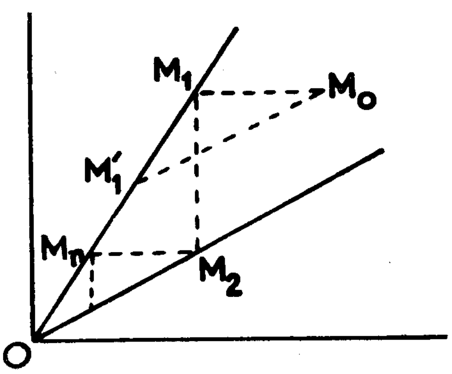
$$\phi_Y(u) = e^{iby\, u} \phi_2(u)$$

$$\text{donc } \phi(u, v) = \phi_1(v)\, \phi(u + av, 0)$$
$$= \phi_2(u)\, \phi(0, v + bu)$$

avec $\phi_1(v) = \dfrac{\phi\,(0, v)}{\phi\,(av, 0)} \qquad \phi_2(u) = \dfrac{\phi(u, 0)}{\phi(0, bu)}$ (4)

donc $\phi(u, v) = \dfrac{\phi(u + av, 0)\,\phi(0, v)}{\phi(av, 0)} = \dfrac{\phi(0, v + bu)\,\phi\,(u, 0)}{\phi(0, bu)}$ (5)

or $\phi(u, v)$ ne peut s'annuler en un point $M_0\,(u_0, v_0)$ sans quoi d'après le second membre de (5) $\phi(u, v)$ doit s'annuler soit sur la droite $u + av = u_0 + a\,v_0$ soit sur la droite $v = v_0$ donc au point M_1 (ou M'_1) de la droite $v + bu$. Le troisième membre de (5) devant aussi s'annuler, il est nécessaire qu'il soit nul soit sur la droite $v + bu = 0$ (ce qui est impossible ϕ devant être égale à 1 à l'origine) soit en M_2. On voit ainsi que ϕ doit s'annuler en une suite de points M_n qui tend vers 0 ce qui est impossible en vertu de la continuité de ϕ à l'origine.



On peut donc prendre les logarithmes des deux membres de (5) et on a $\psi(u, v) = G_1\,(u + av) + \psi_1(v) = H_1(v + bu) + \psi_2(u)$ (6) donnons à u et à v des accroissements h et k tels que $k = -bh$

$G_1(u + av + (1 - ab)h) - G(u + av) = \psi_2(u + h) - \psi_2\,(u) + \psi_1\,(v + k) - \psi_1(v)$ supposons $ab \neq 1$ (sans quoi les variables aléatoires XY sont indépendantes) alors $\triangle G\,(u + av) = \triangle\psi_2\,(u) + \triangle\psi_1(v)$ (7) donc $\triangle G$ est fonction linéaire de $u = av$ et G est un polynome du second degré.

Il s'ensuit que si l'on exclut le cas trivial de l'indépendance, la loi normale est la seule loi pour laquelle on a corrélation dure de X par rapport à y et Y par rapport à X les lignes de régression étant des droites.

Par contre, si on se pose le problème de Bernstein en supposant qu'une version de la fonction de répartition liée est de la forme $F_x(y) = C\lambda'_1(x)\,[y - \alpha(x)]'_1$ on obtient une légère généralisation de la notion de corrélation isogène de Sarmanov et on sait dans ce cas que plusieurs autres surfaces répondent à la question.

Si maintenant nous nous posons un problème concret quelconque, nous constaterons généralement que nous ne savons pas le résoudre sur le couple XY proposé mais qu'on le résoud aisément pour un couple X*Y* obéissant à une loi normale. Les calculs faits dans la pratique aboutissent ainsi à substituer au couple donné initialement un certain couple aléatoire X*Y* qui vérifie les conditions de Bernstein. Ce couple est généralement choisi de manière que les moments des deux premiers ordres de X* Y* soient les mêmes que ceux de XY ce qui détermine entièrement X*Y* chaque fois que X et Y ne sont pas indépendantes.

On peut dire que ce procédé aboutit à prendre pour couple X*Y* un couple dans lequel les lois marginales de X* et Y* ont mêmes moyenne et même variance que ceux de X et Y et dans laquelle les cœfficients de corrélation sont les mêmes. Il est remarquable de constater que la droite de régression de Y* en x est aussi la droite de régression ajustée par la méthode des moindres carrés, la distance « moyenne » du couple aléatoire XY à cette droite représentant l'écart type lié du couple aléatoire X*Y*. Telle est l'idée qui m'a conduit à considérer le couple aléatoire X*Y* classique comme bien défini quand on s'est donné :

1° La loi marginale de X* ;
2° La ligne de régression de Y* en x ;
3° La fonction de répartition liée, abstraction faite du paramètre de position : C.

C'est la théorie que j'ai développée dans Feron (1) sans toutefois mentionner son interprétation géométrique.

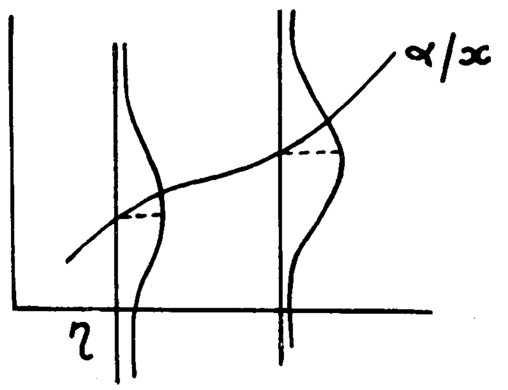
Dans le cas où on choisit ces 3 données de la manière qui a été indiquée précédemment, il se trouve que le couple X*Y* obéit à une loi normale de sorte que X* est aussi en corrélation dure par rapport à y de sorte que ce couple obéit à une surface de Bernstein.

Il n'en va plus de même si nous imaginons d'autres procédés plus logiques pour le choix de 1°, 2°, 3°.

Ceci nous amène à dire que :

a. Un couple aléatoire XY obéit à une loi de S. Bernstein généralisée si il existe une version de la fonction de répartition liée $F_x(y)$ qui reste invariante à une translation près $F_x(y) = C[y - \alpha(x)]$.

Ici on n'impose plus rien à $F_y(x)$. Dans le cas où une densité existe, une « surface de Bernstein généralisée » apparaîtra donc comme engendrée par un fil de fer rigide, représentant la densité de probabilité liée $f_x(y) = c[y - \alpha(x)]$ qui se déplace sans déformation de manière qu'un de ses points se projette sur la courbe $\alpha(x)$;



b. Nous dirons au contraire que le couple XY obéit à une loi de SARMANOV généralisée si la fonction de répartition liée $F_x(y)$ reste invariante à une translation et à une homothétie près. $F_y(y) = C[\lambda(x)\ y - \alpha(x)]$.

Dans le cas où existe une densité, ceci équivaut à supposer que la densité liée peut subir une déformation élastique.

Ce sont les définitions *a* et *b* dont je suis parti dans FERON (1) pour élaborer une nouvelle théorie généralisant les notions de régression et de corrélation.

Il est remarquable de constater que nous pouvons retrouver ainsi les notions classiques de régression ainsi que tous les indices de corrélation importants. Nous retrouvons ainsi le carré du coefficient linéaire r^2, le rapport de corrélation de PEARSON η^2 les indices simples et quadratiques de connexion de Gini. Par contre nous ne retrouvons pas les indices donnant, quand le nombre de cases du tableau tend vers l'infini, des valeurs tendant presque toujours vers 0 ou 1 comme ceux de Jordan, Geiringer, Tschuprow.

Nous pouvons évidemment généraliser à un couple de vecteurs aléatoires les notions de lois de BERNSTEIN et de SARMANOV. On retrouve ainsi notamment les notions de coefficient de corrélation partielle et de coefficient de corrélation multiple.

Il est d'ailleurs bien évident que l'on peut également appliquer les méthodes précédentes à la plupart des problèmes concrets de probabilités ou de statistique. Je pense ici plus particulièrement à la théorie des tests. Dans ce domaine aucun travail dans cette voie n'a été fait à ma connaissance.

Manuscrit reçu le 20 août 1966.

BIBLIOGRAPHIE

S. BERNSTEIN (1). — *Fondements géométriques de la théorie des Corrélations Metron* 1928.

FERON (1). — *Information, régression, corrélation* « Pub. Inst. Stat. Univ ». Paris 1956.

FERON (2). — *Mérites comparés des divers indices de corrélation* « Journ. Soc. Stat. » Paris 88, 328-53 Sept. Oct. 1947.

FERON (3). — *De l'information* « C.R. AC. SC. » 230 ,p. 1945-7 1950.

FERON (4). — *Information et régression* « C.R. AC. SC. » 232, p. 1636-8, 1951.

FERON (5). — *Information et corrélation* « C.R. AC. SC. » 234, p. 1343-5, 1952.

FERON (6). — *Convexité et information* « C.R. AC. SC. » 234, p. 1840-1, 1952.

FERON (7). — *Tableaux de corrélation* « C.R. 35ᵉ Session Inst. Intern. Stat. » 89, 1965.

FERON et FOURGEAUD (1). — *Quelques propriétés caractéristiques de la loi de Laplace Gauss*, « Pub. Inst. Stat. Univ. Paris », Vol. 1, fasc. 3, 1952, p. 21-25.

SARMANOV (1). — *De la création isogène (en russe)* « C.R. AC. SC. » U.R.S.S., 1945.