Bernoulli **7**(2), 2001, 343–350

Remarks on the maximum correlation coefficient

AMIR DEMBO[1], ABRAM KAGAN[2] and LAWRENCE A. SHEPP[3]

[1]*Department of Statistics and Department of Mathematics, Stanford University, STANFORD CA 94305, USA. E-mail: amir@math.stanford.edu*
[2]*Department of Mathematics, University of Maryland, College Park MD 20742-4015, USA. E-mail: amk@math.umd.edu*
[3]*Department of Statistics, 501 Hill Center, Busch Campus, Rutgers University, Piscataway NJ 08855, USA. E-mail: shepp@stat.rutgers.edu*

The maximum correlation coefficient between partial sums of independent and identically distributed random variables with finite second moment equals the classical (Pearson) correlation coefficient between the sums, and thus does not depend on the distribution of the random variables. This result is proved, and relations between the linearity of regression of each of two random variables on the other and the maximum correlation coefficient are discussed.

Keywords: correlation; linear regression; maximum correlation; spherically symmetric distributions; sums of independent random variables

1. Introduction

Let X_1, X_2 be random elements defined on a probability space $(\mathscr{X}, \mathscr{A}, P)$ taking values in $(\mathscr{X}_1, \mathscr{B}_1), (\mathscr{X}_2, \mathscr{B}_2)$, respectively. The map X_i: $(\mathscr{X}, \mathscr{A}) \Rightarrow (\mathscr{X}_i, \mathscr{B}_i)$ generates the subalgebra $\mathscr{A}_i = X_i^{-1}(\mathscr{B}_i)$ of $\mathscr{A}$, $i = 1, 2$. Denote by P_i the restriction of the measure P on $\mathscr{A}_i$, $i = 1, 2$. Let $L^2 = L^2(P)$ be the Hilbert space of $\mathscr{A}$-measurable functions φ with finite $\mathrm{E}|\varphi|^2 = \int |\varphi(x)|^2 \mathrm{d}P$ and inner product $(\varphi_1, \varphi_2) = \mathrm{E}(\varphi_1 \varphi_2)$, and let $L_i^2 = L^2(P_i)$ be the Hilbert space of $\mathscr{A}_i$-measurable functions with finite $\mathrm{E}|\varphi|^2$ and the same inner product. Plainly, L_i^2 is a (closed) subspace of L^2, $i = 1, 2$.

The maximum correlation coefficient (or maximum correlation for short) between X_1 and X_2, introduced in Gebelein (1941), is

$$R(X_1, X_2) = \sup \rho(\varphi_1(X_1), \varphi_2(X_2)), \tag{1}$$

the supremum being taken over all (non-constant) $\varphi_1 \in L_1^2$, $\varphi_2 \in L_2^2$. As usual, $\rho(\xi, \eta)$ denotes the classical (Pearson) correlation between random variables ξ and η. The maximum correlation $R(X_1, X_2)$ vanishes if and only if X_1 and X_2 are independent or, equivalently, if and only if the subspaces L_1^2 and L_2^2 are orthogonal. In general, $R(X_1, X_2)$ is the cosine of the angle between L_1^2 and L_2^2,

$$R(X_1, X_2) = \cos(L_1^2, L_2^2).$$

Czáki and Fisher (1963) studied the maximum correlation as a geometric characteristic.

The following observation is due to Rényi (1959). If

$$R(X_1, X_2) = \rho(\varphi_1, \varphi_2) = R, \tag{2}$$

say, for some φ_i with $\mathrm{E}(\varphi_i) = 0$, $\mathrm{E}(\varphi_i^2) = 1$, $i = 1, 2$, then necessarily

$$\mathrm{E}(\varphi_1 | X_2) = R\varphi_2, \qquad \mathrm{E}(\varphi_2 | X_1) = R\varphi_1. \tag{3}$$

Rényi (1959) also gives sufficient conditions on (X_1, X_2) for (2) to hold with φ_1, φ_2 satisfying (3) for some $R > 0$.

Based on (3), Breiman and Friedman (1985) suggested an alternating conditional expectations algorithm for finding φ_1, φ_2 such that $\rho(\varphi_1, \varphi_2)$ is maximized. They also showed how the maximizing φ_1, φ_2 can be estimated from observations of (X_1, X_2). If (X_1, X_2) is a bivariate Gaussian random vector with $\rho(X_1, X_2) = \rho$, then it has long been known that

$$R(X_1, X_2) = |\rho|. \tag{4}$$

There are several proofs of (4); see, for example, Lancaster (1957).

Now let $Y_1, Y_2 \ldots$ be independent and identically distributed (non-degenerate, i.e. with distribution not concentrated at a point) random variables with $\mathrm{var}(Y_i) < \infty$. Set $S_k = Y_1 + \ldots + Y_k$. We prove in Section 2 that, for $m \leqslant n$,

$$R(S_m, S_n) = \rho(S_m, S_n) = \sqrt{m/n}, \tag{5}$$

and thus $R(S_m, S_n)$ does not depend on the distribution of Y_i. To the best of the authors' knowledge, this result is new. It is a little unexpected given that $R(S_m, S_n)$ is a very nonlinear characteristic of the sums. The special case of (5) with $m = 1$, $n = 2$ was known to Samuel Karlin. His advice on approaching the general case was most apposite.

It is not known if (5) holds when $\mathrm{var}(Y_i) = \infty$. Our arguments only tell us that it is always true that

$$R(S_m, S_n) \leqslant \sqrt{m/n}, \qquad m \leqslant n.$$

The normalized sums

$$\tilde{S}_m = \frac{S_m - \mathrm{E}(S_m)}{\sqrt{\mathrm{var}(S_m)}}, \qquad \tilde{S}_n = \frac{S_n - \mathrm{E}(S_n)}{\sqrt{\mathrm{var}(S_n)}}$$

satisfy condition (3) with $R = \sqrt{m/n}$. However, the sufficient conditions in Rényi (1959) for (3) to imply (2) are not satisfied for $\tilde{S}_m$, $\tilde{S}_n$ constructed from arbitrary $Y_1, \ldots, Y_n$ with $\mathrm{var}(Y_i) < \infty$. Our proof is based on the Efron–Stein (Efron and Stein, 1981) decomposition.

In Section 3, random vectors (X_1, X_2) with

$$\mathrm{E}(X_1 | X_2) = aX_2, \qquad \mathrm{E}(X_2 | X_1) = bX_1 \tag{6}$$

are considered, for some constants a, b. Condition (6) is easily seen to be necessary for

$$R(X_1, X_2) = |\rho(X_1, X_2)| \tag{7}$$

to hold. Indeed, assuming (without loss of generality) that $\mathrm{E}(X_1) = \mathrm{E}(X_2) = 0$, setting $\Lambda_i = \{cX_i,\ c \in \mathbb{R}\}$, $i = 1, 2$, and denoting by $\hat{E}(\cdot|\Lambda)$ the projection operator into the subspace Λ, (6) is equivalent to

$$\hat{E}(X_1|L_2^2) = \hat{E}(X_1|\Lambda_2), \qquad \hat{E}(X_2|L_1^2) = \hat{E}(X_2|\Lambda_1). \tag{8}$$

If the first relation in (8) does not hold then

$$\cos(\Lambda_1,\ L_2^2) > \cos(\Lambda_1,\ \Lambda_2)$$

and a fortiori

$$R(X_1,\ X_2) = \cos(L_1^2,\ L_2^2) > \cos(\Lambda_1,\ \Lambda_2) = |\rho(X_1,\ X_2)|.$$

Remarks in Sarmanov (1958a; 1958b) can be interpreted as saying that (6) is sufficient for (7) – this is the interpretation of Szekely and Gupta (1998). We show in Section 3 that (6) is only necessary for (7).

2. Maximum correlation between sums of independent and identically distributed random variables

Our main tool is an expansion of the analysis of variance type due to Efron and Stein.

Lemma 1. *Let $Y_1, \ldots, Y_k$ be independent and identically distributed random variables. For any symmetric function $h(Y_1, \ldots, Y_k)$ with $\mathrm{E}(h) = 0$, $\mathrm{E}(h^2) < \infty$, the following expansion holds:*

$$\begin{aligned} h(Y_1, \ldots, Y_k) = \sum_{1 \leqslant i_1 \leqslant k} h_1(Y_{i_1}) + \sum_{1 \leqslant i_1 < i_2 \leqslant k} h_2(Y_{i_1},\ Y_{i_2}) \\ + \sum_{1 \leqslant i_1 < i_2 < i_3 \leqslant k} h_3(Y_{i_1},\ Y_{i_2},\ Y_{i_3}) + \ldots + h_k(X_1,\ X_2, \ldots, X_k), \end{aligned} \tag{9}$$

where, for all $j = 1, \ldots, l$ and $l = 1, \ldots, k$,

$$\mathrm{E}(h_l(Y_{i_1}, \ldots, Y_{i_l})|\{Y_{i_1}, \ldots, Y_{i_l}\} \backslash Y_{i_j}) = 0. \tag{10}$$

Proof. See Efron and Stein (1981). □

The orthogonality property (10) implies that the (symmetric zero-mean) function

$$\hat{h}(Y_1, \ldots, Y_l) = \mathrm{E}\{h(Y_1, \ldots, Y_k)|Y_1, \ldots, Y_l\}$$

can be decomposed in the form of (9) and with the same functions $h_1, \ldots, h_l$ as in (9) but with their arguments running over the set $(Y_1, \ldots, Y_l)$:

$$\hat{h}(Y_1, \ldots, Y_l) = \sum_{1\leqslant i_1 \leqslant l} h_1(Y_{i_1}) + \sum_{1\leqslant i_1 < i_2 \leqslant l} h_2(Y_{i_1}, Y_{i_2}) + \ldots + h_l(X_1, \ldots, X_l). \tag{11}$$

For $j > l$,

$$\mathrm{E}\{h_j(Y_{i_1}, \ldots, Y_{i_j})|Y_1, \ldots, Y_l\} = 0$$

since among $Y_{i_1}, \ldots, Y_{i_j}$ there is at least one random variable different from all $Y_1, \ldots, Y_l$. In calculating $\mathrm{E}\{h(Y_1, \ldots, Y_k)\}^2$ using (9), all the cross product terms vanish, since if $r < q$ then

$$\mathrm{E}\{h_r(Y_{i_1}, \ldots, Y_{i_r})h_q(Y_{j_1}, \ldots, Y_{j_q})\} = \mathrm{E}\{h_r(Y_{i_1}, \ldots, Y_{i_r})\mathrm{E}(h_q(Y_{j_1}, \ldots, Y_{j_q})|Y_{i_1}, \ldots, Y_{i_r})\} = 0$$

(among $Y_{j_1}, \ldots, Y_{j_q}$ there is at least one random variable different from all $Y_{i_1}, \ldots, Y_{i_r}$). The same holds for $\mathrm{E}\{\hat{h}(Y_1, \ldots, Y_l)\}^2$.

Having made these remarks, we can state the next lemma.

Lemma 2. *Let* $Y_1, Y_2, \ldots$ *be independent and identically distributed random variables,* $S_k = Y_1 + \ldots = Y_k$. *If* $\mathrm{E}\{h(S_k)\}^2 < \infty$ *then, for* $l \leqslant k$,

$$\mathrm{E}\{\mathrm{E}(h(S_k)|S_l)\}^2 \leqslant (l/k)\mathrm{E}\{h(S_k)\}^2 + (1 - l/k)\{\mathrm{E}(h(S_k))\}^2. \tag{12}$$

Proof. Inequality (12) is a special case of the following inequality holding for any symmetric function $h(Y_1, \ldots, Y_k)$ with $\mathrm{E}(h^2) < \infty$:

$$\mathrm{E}\{\mathrm{E}(h(Y_1, \ldots, Y_k)|Y_1, \ldots, Y_l)\}^2 \leqslant (l/k)\mathrm{E}\{h(Y_1, \ldots, Y_k)\}^2 + (1 - l/k)\{\mathrm{E}(h(Y_1, \ldots, Y_k))\}^2. \tag{13}$$

Indeed, $h(S_k) = h(Y_1 + \ldots + Y_k)$ is symmetric in $Y_1, \ldots, Y_k$. Furthermore, if ξ, η are independent random elements then, for any functions $g(\xi)$, $h(g(\xi), \eta)$ with $\mathrm{E}|h| < \infty$,

$$\mathrm{E}\{h(g(\xi), \eta)|\xi\} = \mathrm{E}\{h(g(\xi), \eta)|g(\xi)\},$$

whence, for $l \leqslant k$,

$$\begin{aligned} \mathrm{E}\{h(S_k)|Y_1, \ldots, Y_l\} &= \mathrm{E}\{h(S_l + Y_{l+1} + \ldots + Y_k)|Y_1, \ldots, Y_l\} \\ &= \mathrm{E}\{h(S_l + Y_{l+1} + \ldots + Y_k)|S_l\} = \mathrm{E}\{h(S_k)|S_l\}. \end{aligned}$$

Thus, (12) follows from (13).

In proving (13), one may always assume $\mathrm{E}\{h(X_1, \ldots, X_k)\} = 0$; then $\mathrm{E}\{\hat{h}(X_1, \ldots, X_l)\} = 0$. By virtue of Lemma 1,

$$\mathrm{E}\{h(Y_1, \ldots, Y_k)\}^2 = \binom{k}{1}\mathrm{E}(h_1^2) + \binom{k}{2}\mathrm{E}(h_2^2) + \ldots + \binom{k}{k}\mathrm{E}(h_k^2) \tag{14}$$

and

$$\mathrm{E}\{\hat{h}(Y_1, \ldots, Y_l)\}^2 = \binom{l}{1}\mathrm{E}(h_1^2) + \binom{l}{2}\mathrm{E}(h_2^2) + \ldots + \binom{l}{l}\mathrm{E}(h_l^2). \tag{15}$$

Noting that, for $1 \leqslant r \leqslant l \leqslant k$,

$$(l/k)\binom{k}{r} = \frac{l}{k}\frac{k(k-1)\dots(k-r+1)}{r!} \geqslant \frac{l(l-1)\dots(l-r+1)}{r!} = \binom{l}{r},$$

whence

$$\begin{aligned}
\mathrm{E}\{\hat{h}(Y_1, \dots, Y_l)\}^2 &\leqslant (l/k)\left\{\binom{k}{1}\mathrm{E}(h_1^2) + \dots + \binom{k}{l}\mathrm{E}(h_l^2)\right\} \\
&\leqslant (l/k)\left\{\binom{k}{1}\mathrm{E}(h_1^2) + \dots + \binom{k}{l}\mathrm{E}(h_l^2)\right. \\
&\quad \left. + \binom{k}{l+1}\mathrm{E}(h_{l+1}^2) + \dots + \binom{k}{k}\mathrm{E}(h_k^2)\right\} \\
&= (l/k)\mathrm{E}\{h(Y_1, \dots, Y_k)\}^2,
\end{aligned}$$

which is exactly (13). □

We now state and prove our main result.

Theorem 1. *Let $Y_1, Y_2, \dots$ be independent and identically distributed non-degenerate random variables with $\mathrm{E}(Y_i^2) < \infty$, $S_k = Y_1 + \dots + Y_k$. The maximum correlation between S_m and S_n equals the (Pearson) correlation, and thus does not depend on the distribution of Y_i:*

$$R(S_m, S_n) = \rho(S_m, S_n) = \sqrt{m/n}, \qquad m \leqslant n. \tag{16}$$

Proof. Take $\varphi_1(S_m)$, $\varphi_2(S_n)$ such that

$$\mathrm{E}\{\varphi_1(S_m)\} = \mathrm{E}\{\varphi_2(S_n)\} = 0, \qquad \mathrm{E}\{\varphi_1(S_m)\}^2 < \infty, \qquad \mathrm{E}\{\varphi_2(S_n)\}^2 < \infty. \tag{17}$$

Then

$$\mathrm{E}\{\varphi_1(S_m)\varphi_2(S_n)\} = \mathrm{E}\{\varphi_1(S_m)\mathrm{E}(\varphi_2(S_n)|S_m)\},$$

and, by the Cauchy–Schwarz inequality,

$$\begin{aligned}
|\mathrm{E}\{\varphi_1(S_m)\varphi_2(S_n)\}|^2 &\leqslant \mathrm{E}\{\varphi_1(S_m)\}^2\mathrm{E}\{\mathrm{E}(\varphi_2(S_n)|S_m)\}^2 \\
&\leqslant (m/n)\mathrm{E}\{\varphi_1(S_m)\}^2\mathrm{E}\{\varphi_2(S_n)\}^2,
\end{aligned} \tag{18}$$

the second inequality in (18) being due to (12).

Since (18) holds for any $\varphi_1(S_m)$, $\varphi_2(S_n)$ subject to (17),

$$R^2(S_m, S_n) \leqslant m/n. \tag{19}$$

On the other hand,

$$\rho(S_m, S_n) = \frac{\mathrm{E}\{(S_m - \mathrm{E}(S_m))(S_n - \mathrm{E}(S_n))\}}{\sqrt{\mathrm{var}(S_m)\mathrm{var}(S_n)}} = \sqrt{\frac{m}{n}},$$

so that

$$R(S_m, S_n) \geqslant \sqrt{m/n}. \tag{20}$$

The last two inequalities imply (16). □

The above arguments also prove that, when $\mathrm{E}(Y_i^2) = \infty$, $R(S_m, S_n) \leqslant \sqrt{m/n}$.

3. Linear regression and maximum correlation

We start with a simple example of non-degenerate random variables X_1, X_2 with

$$\mathrm{E}(X_1|X_2) = \mathrm{E}(X_2|X_1) = 0$$

and

$$R(X_1, X_2) > |\rho(X_1, X_2)| = 0.$$

Let U_1, U_2, W be independent random variables with

$$P(U_i = -1) = P(U_i = 1) = \tfrac{1}{2}, \quad i = 1, 2, \qquad 0 < \mathrm{var}(W) < \infty.$$

Set $X_1 = U_1 W$, $X_2 = U_2 W$. Since

$$\mathrm{E}(X_1|U_2, W) = \mathrm{E}(U_1 W|U_2, W) = W\mathrm{E}(U_1) = 0,$$

then $\mathrm{E}(X_1|X_2) = 0$ and, similarly, $\mathrm{E}(X_2|X_1) = 0$, whence

$$\rho(X_1, X_2) = 0.$$

However, $P(X_1^2 = X_2^2) = 1$, and thus

$$R(X_{1,}X_2) = 1.$$

This example was constructed in response to a question asked by Sid Browne of Columbia University.

A random vector $(U_1, U_2, \ldots, U_n)$ has *spherically symmetric* distribution if

$$f(t_1, t_2, \ldots, t_n) = \mathrm{E}\{\exp \mathrm{i}(t_1 U_1 + t_2 U_2 + \ldots + t_n U_n)\} = g(t_1^2 + t_2^2 + \ldots + t_n^2),$$

for all $t_1, t_2, \ldots, t_n \in \mathbb{R}$. The analytical and statistical properties of spherically symmetric (and, more generally, elliptically contoured) distributions have been studied by many authors – see Fang *et al.* (1990), Gupta and Varga (1993) and references therein.

Assume that the covariance matrix B of $U_1, U_2, \ldots, U_n$ exists. If

$$X_1 = a_1 U_1 + a_2 U_2 + \ldots + a_n U_n, \qquad X_2 = b_1 U_1 + b_2 U_2 + \ldots + b_n U_n \tag{21}$$

are linear forms in $U_1, U_2, \ldots, U_n$ with non-random coefficients, then

$$\mathrm{E}(X_1|X_2) = \lambda_1 X_2, \qquad \mathrm{E}(X_2|X_1) = \lambda_2 X_1 \tag{22}$$

for some λ_1, λ_2 (see Eaton 1986). This means that for *uncorrelated* X_1, X_2,

$$\mathrm{E}(X_1|X_2) = \mathrm{E}(X_2|X_1) = 0.$$

If for all linear forms (21)

$$R(X_1, X_2) = |\rho(X_1, X_2)|,$$

then for all uncorrelated forms X_1, X_2

$$R(X_1, X_2) = 0,$$

i.e.,

$$\text{uncorrelatedness of } X_1, X_2 \text{ implies their independence.} \tag{23}$$

Vershik (1964) showed that if rank $B \geqslant 2$ then (23) is equivalent to the random vector $(U_1, U_2, \ldots, U_n)$ being Gaussian.

Thus, for any non-Gaussian vector $(U_1, U_2, \ldots, U_n)$ with spherically symmetric distribution and covariance matrix of rank $\geqslant 2$, there exists a pair of linear forms (21) with (22) such that

$$R(X_1, X_2) > |\rho(X_1, X_2)|.$$

Note in passing that for bivariate vectors (U_1, U_2) Vershik's result can be slightly modified. According to this modification, if (U_1, U_2) is an arbitrary non-degenerate random vector (with no moment assumption a priori) such that, for any $X_1 = a_1 U_1 + a_2 U_2$, there exists a non-trivial form $X_2 = b_1 U_1 + b_2 U_2$ (i.e., with $b_1^2 + b_2^2 > 0$) independent of X_1, then (U_1, U_2) is Gaussian.

To prove this, take a pair of independent forms X_1, X_2. Plainly they are linearly independent, and thus any linear form in U_1, U_2 is a linear combination of X_1, X_2. Now take $X_1' = a_1' X_1 + a_2' X_2$ with $a_1' a_2' \neq 0$ and find $X_2' = b_1' X_1 + b_2' X_2$ independent of X_1'. Independence of (i) X_1 and X_2 and of (ii) X_1' and X_2' results in $b_1' b_2' \neq 0$. By virtue of the Bernstein–Kac theorem (a very special case of the Darmois–Skitovich theorem; see, for example, Kagan *et al.*, 1973, Chapter 3), X_1 is Gaussian (as is X_2). Since X_1 is arbitrary, the Cramér–Wold principle implies that (U_1, U_2) is a Gaussian vector.

4. References

Breiman, L. and Friedman, J. (1985) Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.*, **80**, 580–619.

Czáki, P. and Fisher, J. (1963) On the general notion of maximum correlation. *Magyar Tudományos Akad. Mat. Kutató Intézetenk Közleményei* (Publ. Math. Inst. Hungar. Acad. Sci.), **8**, 27–51.

Eaton, M. (1986) A characterization of spherical distributions. *J. Multivariate Anal.*, **20**, 272–276.

Efron, B. and Stein, C. (1981) The jackknife estimate of variance. *Ann. Statist.*, **9**, 586–596.

Fang, K.-T., Kotz, S., Ng, K.-W. (1990) *Symmetric Multivariate and Related Distributions*. London: Chapman & Hall.

Gebelein, H. (1941) Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *Z. Angew. Math. Mech.*, **21**, 364–379.

Gupta, A. and Varga, T. (1993) *Elliptically Contoured Models in Statistics*. Dordrecht: Kluwer.
Kagan, A.M., Linnik, Yu.V. and Rao, C.R. (1973) *Characterization Problems in Mathematical Statistics*. Wiley.
Lancaster, H.O. (1957) Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika*, **44**, 289–292.
Rényi, A. (1959) On measures of dependence. *Acta Math. Acad. Sci. Hungar.*, **10**, 441–451.
Sarmanov, O.V. (1958a) Maximum correlation coefficient (symmetric case). *Dokl. Akad. Nauk SSSR*, **120**, 715–718 (in Russian).
Sarmanov, O.V. (1958b) Maximum correlation coefficient (non-symmetric case). *Dokl. Akad. Nauk SSSR*, **121**, 52–55 (in Russian).
Szekely, G.J. and Gupta, A.K. (1998) On a paper of V.B. Nevzorov. *Math. Meth. Statist.*, **2**, 122.
Vershik, A.M. (1964) Some characteristic properties of Gaussian stochastic processes. *Theory Probab. Appl.*, **9**, 353–356 (in Russian).

Received May 2000