**DIMACS Technical Report 2001-25**
**August 2001**

# Remarks on the maximum correlation coefficient

by

Amir Dembo
Stanford University, Stanford, CA, USA

Abram Kagan [1]
University of Maryland, College Park, MD, USA

Lawrence A. Shepp
Rutgers University, Piscataway, N. J., USA

[1]Corresponding author

**ABSTRACT**

The maximum correlation coefficient between partial sums of independent 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. Besides proving this, relations between linearity of regression of each of two random variables on the other and the maximum correlation coefficient are discussed.

*Keywords and phrases*: Correlation, maximum correlation, sums of iid random variables, linear regression, spherically symmetric distributions.
*AMS subject classification* : Primary 62H10, secondary 62J02, 62J05.

Running title: *Maximum correlation coefficient*

# 1 Introduction

Let $X_1, X_2$ be random elements defined on a probability space $(\mathcal{X}, \mathcal{A}, P)$ taking values in $(\mathcal{X}_1, \mathcal{B}_1), (\mathcal{X}_2, \mathcal{B}_2)$, respectively. The map $X_i : (\mathcal{X}, \mathcal{A}) \Rightarrow (\mathcal{X}_i, \mathcal{B}_i)$ generates the subalgebra $\mathcal{A}_i = X_i^{-1}(\mathcal{B}_i)$ of $\mathcal{A}, i = 1, 2$.
Denote by $P_i$ the restriction of the measure $P$ on $\mathcal{A}_i,\ i = 1, 2$.
Let $L^2 = L^2(P)$ be the Hilbert space of $\mathcal{A}$-measurable functions $\varphi$ with finite $E|\varphi|^2 = \int |\varphi(x)|^2 dP$ and the inner product $(\varphi_1, \varphi_2) = E(\varphi_1\varphi_2)$and let $L_i^2 = L^2(P_i)$ be the Hilbert space of $\mathcal{A}_i$-measurable functions with finite $E|\varphi|^2$ and the same inner product. Plainly, $L_i^2$ is a (closed) subspace of $L^2,\ i = 1, 2$.
The maximum correlation coefficient (shortly, maximum correlation) 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 (nonconstant) $\varphi_1 \in L_1^2$, $\varphi_2 \in L_2^2$. As usual, $\rho(\xi, \eta)$ denotes the classical (Pearson) correlation between random variables $\xi$ and $\eta$.
The maximum correlation $R(X_1, X_2)$ vanishes iff $X_1$ and $X_2$ are independent or, equivalently, iff 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 $L_2^2$,

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

Czáki and Fisher (1963) studied the maximum correlation as a geometric characteristic.
The following observation is due to Rényi (1959). If

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

for some $\varphi_i$ with $E\varphi_i = 0,\ E\varphi_i^2 = 1,\ i = 1, 2$ then necessarily

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

Rényi (1959) also gives sufficient conditions on $(X_1, X_2)$ for (2) to hold with $\varphi_1, \varphi_2$ satisfying (3) with some $R > 0$.
Based on (3), Breiman and Friedman (1985) suggested an alternating conditional expectations algorithm (ACE) for finding $\varphi_1, \varphi_2$ maximizing $\rho(\varphi_1, \varphi_2)$. They also showed how the maximizing $\varphi_1, \varphi_2$ can be estiamted 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 been known since long ago that

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

(there are several proofs of (4); see, for example, Lancaster (1957).)
Let now $Y_1, Y_2 \ldots$ be independent identically distributed (nondegenerate) random variables with $\mathrm{var}(Y_i) < \infty$. Set $S_k = Y_1 + \ldots + Y_k$. We prove in Section 2 that for $m \le n$

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

and, thus, $R(S_m, S_n)$ does not depend on the distribution of $Y_i$. To the best of the authors' knowledge, the 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 $m = 1, n = 2$ had been known to Samuel Karlin. His advise on how to approach the general case was right to the point.
It is not known if (5) holds when $\mathrm{var}(Y_i) = \infty$. Our arguments give only that always

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

The normalized sums

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

satisfy the condition (3) with $R = \sqrt{m/n}$. However, the sufficient conditions in Rényi for (3) to imply (2) are not satisfied for $\tilde{S}_m, \tilde{S}_n$ constructed from arbitrary $Y_1, \dots, Y_n$ with $\mathrm{var}(Y_i) < \infty$. Our proof is based on the Efron-Stein decomposition (Efron and Stein (1981).)
In Section 3, random vectors $(X_1, X_2)$ with

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

for some constants $a, b$ are considered. The condition (6) is easily seen necessary for

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

to hold. Indeed, assuming (without loss in generality) $E(X_1) = E(X_2) = 0$, 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 $\Lambda$, (6) is equivalent to

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

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

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

and a fortiori

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

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

## 2 Maximum correlation between sums of iid random variables

Our main tool is the ANOVA expansion due to Efron and Stein.

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

$$h(Y_1, \dots, Y_k) = \sum_{1 \le i_1 \le k} h_1(Y_{i_1}) + \sum_{1 \le i_1 < i_2 \le k} h_2(Y_{i_1}, Y_{i_2}) + \sum_{1 \le i_1 < i_2 < i_3 \le k} h_3(Y_{i_1}, Y_{i_2}, Y_{i_3}) + \dots + h_k(X_1, X_2, \dots, X_k) \qquad (9)$$

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

$$E(h_l(Y_{i_1}, \dots, Y_{i_l}) | \{Y_{i_1}, \dots, Y_{i_l}\} \setminus Y_{i_j}) = 0. \qquad (10)$$

*Proof.* See Efron, Stein (1981). ♣

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

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

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

$$\hat{h}(Y_1, \dots, Y_l) = \sum_{1 \le i_1 \le l} h_1(Y_{i_1}) + \sum_{1 \le i_1 < i_2 \le l} h_2(Y_{i_1}, Y_{i_2}) + \dots + h_l(X_1, \dots, X_l) \qquad (11)$$

For $j > l$,

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

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

$$\begin{aligned} & E\{h_r(Y_{i_1}, \dots, Y_{i_r}) h_q(Y_{j_1}, \dots, Y_{j_q})\} \\ = \; & E\{h_r(Y_{i_1}, \dots, Y_{i_r}) E(h_q(Y_{j_1}, \dots, Y_{j_q}) | Y_{i_1}, \dots, Y_{i_r})\} = 0 \end{aligned}$$

(among $Y_{j_1}, \dots, Y_{j_q}$ there is at least one random variable different from all $Y_{i_1}, \dots, Y_{i_r}$).
The same holds for $E\{\hat{h}(Y_1, \dots, Y_l)\}^2$.
After these remarks, we can state the next lemma.

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

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

*Proof.* The inequality (12) is a special case of the following inequality holding for any symmetric function $h(Y_1, \dots, Y_k)$ with $E(h^2) < \infty$:

$$E\{E(h(Y_1, \dots, Y_k)|Y_1, \dots, Y_l)\}^2 \le (l/k) E\{h(Y_1, \dots, Y_k)\}^2 + (1 - l/k)\{E(h(Y_1, \dots, Y_k))\}^2. \qquad (13)$$

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

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

whence for $l \leq k$

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

Thus, (12) follows from (13).
In proving (13) one may always assume $E\{h(X_1, \ldots, X_k)\} = 0$;
then $E\{\hat{h}(X_1, \ldots, X_l)\} = 0$. By virtue of Lemma 1,

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

and

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

Noting that for $1 \leq r \leq l \leq k$

$$(l/k)\binom{k}{r} = \frac{l}{k}\frac{k(k-1)\ldots(k-r+1)}{r!} \geq \frac{l(l-1)\ldots(l-r+1)}{r!} = \binom{l}{r}$$

whence

$$\begin{aligned} E\{\hat{h}(Y_1, \ldots, Y_l)\}^2 \leq (l/k)\{\binom{k}{1} E(h_1^2) + \ldots + \binom{k}{l} E(h_l^2)\} \leq \\ (l/k)\{\binom{k}{1} E(h_1^2) + \ldots + \binom{k}{l} E(h_l^2) + \binom{k}{l+1} E(h_{l+1}^2) + \ldots + \binom{k}{k} E(h_k^2)\} = \\ (l/k)E\{h(Y_1, \ldots, Y_k)\}^2 \end{aligned}$$

which is exactly (13). ♣
Here is our main result.

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

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

*Proof of Theorem.* Take $\varphi_1(S_m), \varphi_2(S_n)$ with

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

Then

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

and by Cauchy-Schwarz inequality

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

the last inequality in (18) due to (12).
Since (18) holds for any $\varphi_1(S_m), \varphi_2(S_n)$ subject to (17),

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

On the other side,

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

so that

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

The last two inequalities imply (16).♣
The above arguments also prove that in case of $E(Y_i^2) = \infty$,
$R(S_m, S_n) \leq \sqrt{m/n}$.

# 3 Linear regression and maximum correlation

We start with a simple example of nondegenerate random variables $X_1, X_2$ with

$$E(X_1|X_2) = E(X_2|X_1) = 0$$

and

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

Let $U_1, U_2, W$ be independent random variables with

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

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

$$E(X_1|U_2, W) = E(U_1 W|U_2, W) = W E(U_1) = 0$$

one has $E(X_1|X_2) = 0$ and, similarly, $E(X_2|X_1) = 0$ 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

$$f(t_1, t_2, \ldots, t_n) = E\{\exp i(t_1U_1 + t_2U_2 + \ldots + t_nU_n)\} = g(t_1^2 + t_2^2 + \ldots + t_n^2),$$

for all $t_1, t_2, \ldots, t_n \in \mathbb{R}$. Analytical and statistical properties of spherically symmetric (and, mor generally, elliptically contoured) distributions were studied by many authors (see Fang, Kotz and Ng (1990), Gupta and Varga (1993) and references therein.)
Assume that the covariance matrix $B$ of $U_1, U_2, \ldots, U_n$ exists.
If

$$X_1 = a_1U_1 + a_2U_2 + \ldots + a_nU_n, \quad X_2 = b_1U_1 + b_2U_2 + \ldots + b_nU_n \tag{21}$$

are linear forms in $U_1, U_2, \ldots, U_n$ with nonrandom coefficients then

$$E(X_1|X_2) = \lambda_1X_2,\ E(X_2|X_1) = \lambda_2X_1 \tag{22}$$

for some $\lambda_1, \lambda_2$ (see Eaton (1986).)
It means that for *uncorrelated* $X_1, X_2$

$$E(X_1|X_2) = E(X_2|X_1) = 0.$$

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.,

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

In Vershik (1964) was shown that if rank $B \geq 2$ then the property (23) is equivalent to that the random vector $(U_1, U_2, \ldots, U_n)$ is Gaussian.
Thus, for any non-Gaussian vector $(U_1, U_2, \ldots, U_n)$ with spherically symmetric distribution and covariance matrix of rank $\geq 2$, 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 modified a little. Namely, *if* $U_1, U_2$ *is an arbitrary random vector (with no moment assumption a priori) such that for any* $X_1 = a_1U_1 + a_2U_2$ *there exists a nontrivial form* $X_2 = b_1U_1 + b_2U_2$ *(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 (ii) of $X_1'$ and $X_2'$ results in that $b_1'b_2' \neq 0$. By virtue of Bernstein-Kac theorem (a very special case of the Darmois-Skitovich theorem; see, for example,
Kagan, Linnik, Rao (1973), Ch. 3), $X_1$ (and $X_2$) is Gaussian. 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. Stat. Assoc.*, **80**, 580-619.

Czáki, P. and Fisher, J. (1963). On the general notion of maximum correlation. *Magyar Tudomànyos Akademia Matematikai Kutato Intezetenk Kozlemenyei* (Publication of the Mathematical Institute of the Hungarian Acad. of 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.* Chapman and Hall.

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

Gupta, A., Varga, T. (1993). *Elliptically Contoured Models in Statistics.* Kluwer.

Kagan, A. M., Linnik, Yu. V., 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 Mathematica Academiae Scieniarum Hungaricae*, **10**, 441-451.

Sarmanov, O. V. (1958a). Maximum correlation coefficient (symmetric case). *Doklady Akad. Nauk SSSR*, **120**, 715-718 (in Russian).

Sarmanov, O. V. (1958b). Maximum correlation coefficient (nonsymmetric case). *Doklady Akad. Nauk SSSR*, **121**, 52-55 (in Russian).

Szekely, G. J., Gupta, A. K. (1998). On a paper of V. B. Nevzorov. *Math. Methods Statist.*, **2**, 122.

Vershik, A. M. (1964). Some characteristic properties of Gaussian stochastic processes. *Theory Probab. Applic.*, **9**, 353-356 (in Russian).