The Annals of Statistics
1979, Vol. 7, No. 3, 607–614

SCREENING AND MONOTONIC DEPENDENCE FUNCTIONS IN THE MULTIVARIATE CASE

BY T. KOWALCZYK, A. KOWALSKI, A. MATUSZEWSKI AND E. PLESZCZYŃSKA

Polish Academy of Sciences

An approach to simultaneous treatment of dependence and screening problems is presented. New characterizations of dependence of a random variable X on a random vector Y are obtained by functions $\nu_{X,Y} : (0, 1) \to [0, 1]$ and $\mu_{X,Y} : (0, 1) \to [-1, 1]$ called respectively screening and monotonic dependence functions. These functions are shown to be appropriate measures of the intensity of connection and concordance of X on Y, respectively. The interrelations of ν and μ and their relations to the multiple correlation ratio and the multiple correlation coefficient are demonstrated and illustrated by several examples.

1. Introduction. Statistical literature concerning screening is wildly scattered among many sources and even the terminology is unstable. For instance, Birnbaum (1950a), (1950b) and Marshall and Olkin (1968) considered screening decision schemes on a highly theoretical level as opposed to practical algorithms discussed in statistical quality control textbooks or in psychological papers on occupational and scholastic selection. Roughly speaking, theoretical literature is not very useful for applications, since too much a priori information is needed; on the other hand, applicational papers deal usually with particular examples only and suffer from insufficient theoretical justification.

Surprisingly enough, relations between screening and measures of dependence have not been formalized yet in a clear way. This paper is an attempt to provide one possible formalization by using a general idea of screening to construct some measures of dependence between a random variable X and a random vector Y. It is convenient to keep in mind an interpretation of X as an unobservable "performance" variable and an interpretation of Y as a set of observable "test" variables supplying information about X. Screening is meant here as an operation under which items of some considered population, characterized by values (x, y) of (X, Y), are rejected or accepted on the basis of y's in such a way that the expectation of the performance variable in the population of accepted items should possibly be increased. The fraction of rejected items is assumed to be equal to a preassigned value $p \in (0, 1)$. It seems reasonable to characterize the intensity of dependence of X on Y by means of a suitably normalized expectation of performance variable under best possible screening procedures, for any $p \in (0, 1)$. This

Received June 1977; revised February 1978.

AMS 1970 *subject classifications.* Primary 62G99; Secondary 62H30, 62H20.

Key words and phrases. Screening, measures of monotonic dependence, multiple correlation coefficient, selection, regression functions, truncation.

leads to a functional characterization of dependence called the screening dependence function. On the other hand, the restriction of the set of screening procedures to those assuming some monotonic dependence of X on Y leads to a functional characterization of monotonic dependence called the monotonic dependence function. The latter generalizes to the multivariate case the functional measure of monotonic dependence of X on a random variable Y, first introduced in Kowalczyk and Pleszczyńska (1977) and Kowalczyk (1977).

The concepts introduced provide a new interpretation of traditional real-valued measures of dependence, the multiple correlation ratio and the multiple correlation coefficient, which is given for suitably chosen families of (X, Y) in terms of the quality of the optimal screening. These facts imply that the functions introduced have a potential significance for applications. Moreover, the functions μ and ν will be useful in practice, since qualitative information available in real problems can often by easily expressed in the form of assumptions concerning the shape of μ and ν. In particular, nonlinear models in which ν does not reduce to the multiple correlation coefficient can be considered in various screening problems similarly as it is done in the case of linear models (cf. e.g., Owen and Yueh-ling Hsiao Su (1977)).

2. Definition of screening dependence function. In quality control schemes the term screening is usually applied to finite populations of objects and means a procedure of rejecting a fraction p of items in order to get a more desirable truncated population. In many cases the fraction p is stated in advance as the rate of items admitted to be lost at the price of improving the initial population or as the rate of items for which no reservations are made in the truncated population (admission to educational institutions, personal selection, etc.) Turning from finite populations to probability distributions of random vectors (usually representing selected features of objects in not necessarily finite populations) one has to deal with a general notion of truncated distribution introduced below (cf. e.g., Birnbaum (1950a)).

Let C_n be the class of random vectors $(X, Y = (Y_1, \cdots, Y_n))$ such that $E(X)$ is finite, the distribution of X is nondegenerate and there exists the generalized density f of the distribution P of (X, Y). Let T be any measurable function from R^{n+1} into R and t_p a pth quantile of $T(X, Y)$ for any $p \in (0, 1)$. Moreover, let $\varphi_T : R^{n+1} \to [0, 1]$ be given by the formula

$$\begin{aligned} \varphi_T(x, y) &= 0 \quad \text{if} \quad T(x, y) < t_p, \\ &= \gamma \quad \text{if} \quad T(x, y) = t_p, \\ &= 1 \quad \text{if} \quad T(x, y) > t_p, \end{aligned}$$

where

$$\gamma = \big(1 - p - P(T(X, Y) > t_p)\big)/P(T(X, Y) = t_p) \quad \text{if} \quad P(T(X, Y) = t_p) > 0,$$
$$= 0 \quad \text{otherwise.}$$

It follows from the above definition that for any given p the functions φ_T corresponding to different pth quantiles of $T(X, Y)$ differ on a set of P measure zero. The pair (x, y) is said to be rejected if $T(x, y) < t_p$ and accepted if $T(x, y) > t_p$ while acceptance is randomized with probability γ when $T(x, y) = t_p$. This means that acceptance corresponds to large values of $T(X, Y)$. The distribution with the generalized density given by $f(x, y)\varphi_T(x, y)/(1 - p)$ is referred to as truncated according to T. Obviously, the rejection rate $1 - E(\varphi_T(X, Y))$ is equal to p. The expectation of the first component in the truncated distribution exists for any $(X, Y) \in C_n$ and it will be convenient to denote it by $E(X_{p, T(X, Y)})$.

We concentrate on the situation when one is interested in getting $E(X_{p, T(X, Y)})$ as large as possible. Obviously, the largest value is obtained for T given by $T(x, y) = x$ for any (x, y). Note that the largest value denoted by $E(X_{p, X})$ is equal to the expectation of X under the condition that $X > x_p$ for any pth quantile x_p of X whenever the distribution function of X is continuous in x_p.

If X is an unobservable performance variable, then the optimal screening based directly on x's is not admissible and should be replaced by optimal screening based on y's only. In view of the generalized Neyman-Pearson lemma this is realized by T defined by $T(x, y) = h(y)$, where h is the regression function of X on Y.

Consequently for any $n \geqslant 1$ and any $(X, Y) \in C_n$ it is natural to introduce the expression

$$\nu_{X, Y}(p) = \big(E(X_{p, h(Y)}) - E(X)\big)/\big(E(X_{p, X}) - E(X)\big) \tag{2.1}$$

as a measure of goodness of optimal screening based on y's under the rejection rate p. Obviously, for any $p \in (0, 1)\nu_{X, Y}(p)$ is nonnegative since screening based on the regression function is at least as good as screening according to any constant function T while in the latter case $\varphi_T(x, y) \equiv 1 - p$ and $E(X_{p, T(X, Y)}) = E(X)$.

Given $n \geqslant 1$ and $(X, Y) \in C_n$, the function $\nu_{X, Y}$ defined on $(0, 1)$ by (2.1) will be called the screening dependence function of X on Y.

In view of (2.1), the expectation of the performance variable under optimal screening based on y's when the rejection rate is p is a convex linear combination, with the coefficients $\nu_{X, Y}(p)$ and $1 - \nu_{X, Y}(p)$, of the expectation under optimal screening based on x's and of the expectation when no screening is performed. This expresses the meaning of the screening dependence function in screening problems.

3. Properties of screening dependence function. We shall recall first the definition of the bivariate monotonic dependence function $\mu_{X, Y} : (0, 1) \to [-1, 1]$ given in Kowalczyk and Pleszczyńska (1977) and Kowalczyk (1977): for any $(X, Y) \in C_1$ and $p \in (0, 1)$

$$\begin{aligned} \mu_{X, Y}(p) &= \mu^+_{X, Y}(p) \quad \text{if} \quad \mu^+_{X, Y}(p) > \mu^+_{-X, Y}(p) \\ &= -\mu^+_{-X, Y}(p) \quad \text{otherwise,} \end{aligned} \tag{3.1}$$

where

$$\mu^+_{X, Y}(p) = \big(E(X_{p, Y}) - E(X)\big)/\big(E(X_{p, X}) - E(X)\big). \tag{3.2}$$

This function was shown to indicate the type of monotonic (i.e., positive or negative) dependence of X on Y and to measure its strength. The following theorem establishes the connection between the screening dependence function in the multivariate case and some bivariate monotonic dependence function.

THEOREM 1. *For any* $(X, Y) \in C_n$ $(n = 1, 2, \cdots)$

$$\nu_{X,Y} = \mu_{X,h(Y)},$$

where h *is the regression function of* X *on* Y.

PROOF. By (2.1) and (3.2) $\nu_{X,Y} = \mu^+_{X,h(Y)}$. Then in view of (3.1) the proof is completed by noting that $\nu_{X,Y}$ is nonnegative.

In view of Theorem 1 some further properties of screening dependence functions could be derived from properties of the bivariate monotonic dependence functions.

Throughout this paper we fix the notation h for the regression function of X on Y and G_n for the set of all measurable functions from R^n to R.

THEOREM 2. *For any* $(X, Y) \in C_n (n = 1, 2, \cdots)$

(i) $(\forall_p \in (0, 1))$ $0 \leqslant \nu_{X,Y}(p) \leqslant 1$;

(ii) $\nu_{X,Y}(p) \equiv 0$ iff $h(Y) = EX$ a.s.

(iii) $\nu_{X,Y}(p) \equiv 1$ iff $X = g(Y)$ a.s. *for some* $g \in G_n$;

(iv) *if* $a, b \in R$, $a \neq 0$ *and* f *is one-to-one function* $R^n \to R^n$ *then*

$$\nu_{aX+b,f(Y)}(p) \equiv \nu_{X,Y}(p) \quad \text{if} \quad a > 0,$$

$$\equiv \nu_{X,Y}(1 - p) \quad \text{if} \quad a < 0.$$

PROOF. (i) The statement follows immediately from (2.1) and the considerations in Section 2 following (2.1).

(ii) It follows from Theorem 1 above and Theorem 2.1 (v) in Kowalczyk (1977), since, obviously, the conditional expectation of X given $h(Y)$ is equal to $h(Y)$.

(iii) $\nu_{X,Y}(p) \equiv 1 \Leftrightarrow \mu_{X,h(Y)} \equiv 1$. In view of Theorem 2.1 (v) in Kowalczyk (1977) $\mu_{X,h(Y)}(p) \equiv 1$ is equivalent to the existence of a nondecreasing function $f : R \to R$ such that $X = f \circ h(Y)$ a.s. Then there exists a function $g : R^n \to R$ such that $X = g(Y)$ a.s. It remains to show the reverse implication. Let $X = g(Y)$ a.s., then $g = h$ and, consequently, $\mu^+_{X,h(Y)} \equiv 1$, which completes the proof.

(iv) In view of (2.1) the thesis is true for $f(Y) = Y$. To prove that $\nu_{X,Y}$ is invariant under one-to-one transformations of Y, in view of Theorem 1 it is enough to notice that the conditional expectation of X given $f(Y)$ is equal to $h(Y)$ a.s.

The next theorem describes the conditions under which screening dependence functions are constant and reduce to the multiple correlation coefficient $\rho_{X,Y}$ and the multiple correlation ratio $\eta_{X,Y}$ defined as the correlation coefficient $\rho_{X,h(Y)}$.

THEOREM 3. *Suppose that* $(X, Y) \in C_n$ *and there exists an increasing* $f : R \to R$ *such that* X *and* $f \circ h(Y)$ *have the same distributions. Then*

(i) $(\exists \rho \in (0, 1)\ \nu_{X,Y}(p) \equiv \rho)$ iff f *is linear.*

(ii) *If* $\eta_{X,Y}$ *and* $\rho_{X,Y}$ *exist then*

$$\nu_{X,Y}(p) \equiv \eta_{X,Y} \quad \text{iff } f \text{ is linear,}$$

$$\equiv \rho_{X,Y} \quad \text{iff } f \text{ is linear and } h \text{ is linear in each component.}$$

PROOF. (i) Since the conditional expection of X given $h(Y)$ is equal to $h(Y)$ a.s. then the proof follows immediately from Theorem 1 above and Theorem 2.2 in Kowalczyk (1977).

(ii) The first equivalence follows from (i) and the second part of Theorem 2.2 in Kowalczyk (1977). For the proof of the second equivalence it is enough to recall that, for $\rho_{X,Y} > 0$, $\rho_{X,Y} = \eta_{X,Y}$ iff h is linear in each coordinate.

4. Multivariate monotonic dependence function. Up to now a monotonic dependence function was defined for two random variables, while a screening dependence function was defined for a random variable and a random vector. We shall consider now a straightforward generalization of $\mu_{X,Y}$ for $(X, Y) \in C_n$, $n > 1$. Let G_n^+ be the set of all measurable functions from R^n to R which increase in each coordinate. We define for any $(X, Y) \in C_n$ $(n \geqslant 1)$ and $p \in (0, 1)$

$$\mu^+_{X,Y}(p) = \left(\sup_{g \in G_n^+} E(X_{p,g(Y)}) - E(X)\right) / \left(E(X_{p,X}) - E(X)\right). \tag{4.1}$$

Obviously, the right-hand side of (4.1) reduces for $n = 1$ to $\mu^+_{X,Y}(p)$ given by (3.2) and therefore the same symbol is used in (4.1) and (3.2). Then for $(X, Y) \in C_n$ $(n \geqslant 1)\mu_{X,Y}$ given by (3.1) with $\mu^+_{X,Y}$ defined by (4.1) will be called the monotonic dependence function in the multivariate case. It will serve as a measure of "concordance" while $\nu_{X,Y}$ is a measure of "connection", the notions of concordance and connection being meant here as in Kruskal (1958).

THEOREM 4. *For any* $(X, Y) \in C_n$

(i) $-\nu_{-X,Y} \leqslant \mu_{X,Y} \leqslant \nu_{X,Y}$;

(ii) *if* h *is increasing in each coordinate then* $\nu_{X,Y} = \mu_{X,Y}$;

(iii) $\mu_{X,Y}(p) \equiv 0$ *if* $h(Y) = E(X)$ a.s.;

(iv) *if there exists* $g \in G_n^+$ *such that* $\mu_{X,Y} = \mu_{X,g(Y)}$ *then* $\mu_{X,Y}(p) \equiv 1(-1)$ iff *there exists a nondecreasing (nonincreasing)* $f : R \to R$ *such that* $X = f \circ g(Y)$;

(v) *if* $a, b \in R$, $a \neq 0$ *and* $f_1, \cdots, f_n$ *are one-to-one functions from* R *to* R *then for any* $p \in (0, 1)$

$$\mu_{aX+b,(f_1(Y_1),\cdots,f_n(Y_n))}(p)$$

$$= (\operatorname{sgn} a)\mu_{X,Y}(p) \text{ if } f_1, \cdots, f_n \text{ are increasing,}$$

$$= (-\operatorname{sgn} a)\mu_{X,Y}(1-p) \text{ if } f_1, \cdots, f_n \text{ are decreasing.}$$

PROOF. (i)–(iv) follow immediately from the definitions of μ and ν, Theorem 2 above and Theorems 2.1 and 2.2 in Kowalczyk (1977).

(v) By the definition of $\mu_{X,Y}(p)$ the statement is true for $f_i(Y_i) = Y_i$, $i = 1, \cdots, n$. So it is enough to present the proof for $a = 1$ and $b = 0$. In case of the

first equality, it suffices to notice that for increasing f's

$$\forall g \in G_n^+ \exists g^* \in G_n^+ \forall y = (y_1, \cdots, y_n)\, g(Y) = g^*(f_1(y_1), \cdots, f_n(y_n)).$$

The second equality follows from the first one.

5. Examples.

EXAMPLE 1. Let $(X, Y) \in C_n$ be multinormally distributed. Then by Theorem 3 (ii) and Theorem 2 (ii) $\nu_{X,Y}(p) \equiv \rho_{X,Y}$. Moreover, if the regression coefficients of X on Y are positive then by Theorem 4 (ii) $\mu_{X,Y}(p) \equiv \rho_{X,Y}$. For instance, if $n = 2$ and $r_{X,Y_1} = r_{X,Y_2} = r_{Y_1,Y_2} = 0.5$ then $\nu_{X,Y}(p) \equiv \mu_{X,Y}(p) \equiv 0.58$.

EXAMPLE 2. Let us consider a multivariate t-Student distribution with location vector m, precision matrix T and k degrees of freedom, as introduced in De Groot (1970). If we assume that $k > 1$, we have $(X, Y) \in C_n$. It follows that h is linear and there exist $\alpha > 0$ and β such that $\alpha h(Y) + \beta$ and X have identical distributions. Then by Theorem 3 $\nu_{X,Y}$ is constant and for $k > 2$ it is identically equal to $\rho_{X,Y}$; for $k = 2$ the constant value of $\nu_{X,Y}$ could be treated as a generalization of nonexisting $\rho_{X,Y}$. It is easily seen that if $T = \Sigma^{-1}$ and $(X', Y') \in C_n$ is normally distributed with any mean vector and with covariance matrix Σ then

$$\nu_{X,Y}(p) \equiv \rho_{X',Y'} \equiv \nu_{X',Y'}(p).$$

EXAMPLE 3. Let $(X, Y) \in C_2$ be log-normal, derived from normally-distributed (X', Y') by putting $X = \exp X'$, $Y_i = \exp Y_i'$, $i = 1, 2$. We assume that X_1', Y_1' and Y_2' have zero means and unit variances. The graph in Figure 1 shows the shape of $\nu_{X,Y}$ derived for the case:

$$r_{X',Y_1'} = r_{X',Y_2'} = r_{Y_1',Y_2'} = 0.5. \tag{5.1}$$

The exact formula for $\nu_{X,Y}$ is given in Nalbach-Leniewska (1977). Therein it was proved that $h(Y) = \gamma \cdot \exp(h'(Y'))$, where γ is a positive constant and h' is the regression function of X' on Y' and hence under (5.1) $h(Y)$ is increasing in Y_1 and Y_2; consequently, $\nu_{X,Y} = \mu_{X,Y}$.

Screening and monotonic dependence functions of X on Y are compared in Figure 1 with those of X on Y_1 provided that (5.1) holds. The formula for μ_{X,Y_1} was given in Kowalczyk (1977) while the equality $\nu_{X,Y_1} = \mu_{X,Y_1}$ under (5.1) follows from Theorem 4 (ii). It may also be of interest to compare $\nu_{X,Y}$ and ν_{X,Y_1} with traditional real-valued concepts of dependence. Under (5.1) we have $r_{X,h(Y)} = 0.48$, $\rho_{X,Y} = 0.46$, $r_{X,h_1(Y_1)} = 0.41$ and $r_{X,Y_1} = 0.38$, h_1 being the regression function of X on Y_1.

EXAMPLE 4. Let $(X, Y) \in C_1$ be such that the distribution of Y is symmetric with respect to some $a \in R$ and $X = f(Y)$ a.s., where f is continuous, symmetric with respect to a and decreasing for $y \leqslant a$. Then $\nu_{X,Y}(p) \equiv 1$; $\mu_{X,Y}(p) = -\mu_{X,Y}(1-p)$ for $p \in (0, 1)$ and $\mu_{X,Y}(p)$ is negative for $p \in (0, \frac{1}{2})$; $\nu_{Y,X}(p) \equiv \mu_{Y,X}(p) \equiv$

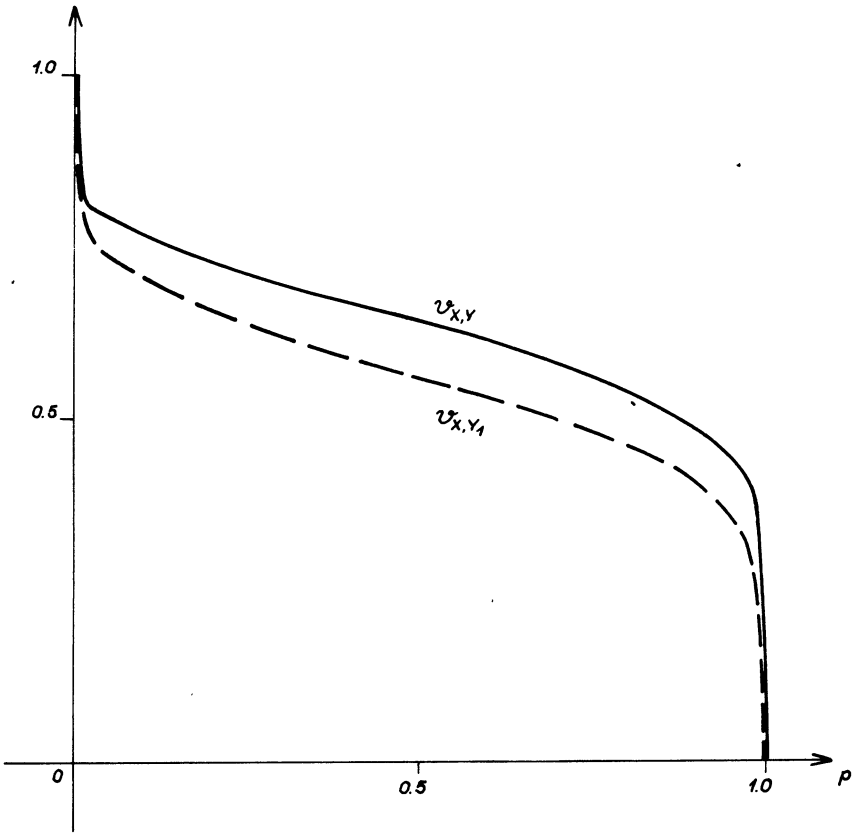

FIG. 1. Screening dependence functions $\nu_{X,Y}$ and ν_{X,Y_1} in the case of bilognormal distribution with correlation coefficients all equal to 0.5 (cf. Example 3).

$\eta_{X,Y} = \rho_{X,Y} = 0$. Note that the shape of $\mu_{X,Y}$ provides a good description of the lack of concordance of X on Y expressed by the change of sign of $\mu_{X,Y}$.

Acknowledgment. We would like to thank the referee for his helpful comments and suggestions.

REFERENCES

[1] BIRNBAUM, Z. W. (1950a). Effect of linear truncation on a multinormal population. *Ann. Math. Statist.* **21** 272–279.

[2] BIRNBAUM, Z. W. (1950b). On optimum selections from multinormal populations. *Ann. Math. Statist.* **21** 443–447.

[3] DE GROOT, M. M. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

[4] KOWALCZYK, T. (1977). General definition and sample counterparts of monotonic dependence functions in the bivariate case. *Math. Operationsforsch. Statist. Series Statist.* **8** 351–369.

[5] KOWALCZYK, T. and PLESZCZYŃSKA, E. (1977). Monotonic dependence functions of bivariate distributions. *Ann. Statist.* **5** 1221–1227.
[6] KRUSKAL, W. (1958). Ordinal measures of association. *J. Amer. Statist. Assoc.* **53** 814–861.
[7] LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. John Wiley, New York.
[8] MARSHALL, A. W. and OLKIN, I. (1968). A general approach to some screening and classification problems. *J. Roy. Statist. Soc. Ser. B* **30** 407–444.
[9] NALBACH-LENIEWSKA, A. (1977). Pattern of dependence of multi-lognormal distribution. Unpublished manuscript.
[10] OWEN, D. B. and YUEH-LING HSIAO, SU (1977). Screening based on normal variables. *Technometrics* **19** 65–68.

INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES
00–901 WARSZAWA PKiN, BOX 22,
POLAND