ESTIMATION BY THE MINIMUM DISTANCE METHOD IN NONPARAMETRIC STOCHASTIC DIFFERENCE EQUATIONS[1]

By J. Wolfowitz

Cornell University

1. Introduction. The present paper is intended to report some of the ideas described in a special invited address delivered by the author at the meeting of the Institute of Mathematical Statistics at Chicago on December 29, 1952. This address dealt with two topics: a) the connection between the method of maximum likelihood and the Wald theory of decision functions, with an explanation of the asymptotic efficiency of the former; and b) estimation by the minimum distance method. The first of these topics is discussed in [1], and this paper will be devoted to a discussion of the second.

The origin of the minimum distance method is to be found in [2]. Applications of the method were extended and generalized in [3]. The paper [4] contains a theorem which is an essential tool. A paper by Kac, Kiefer, and the present author, entitled "On tests of normality and other tests of goodness of fit based on the minimum distance method," is in preparation.

The method of estimation to which this paper is devoted is characterized by the fact that the estimators are always such as to minimize the distance between suitably chosen distribution functions (d.f.). In a variety of problems, which includes many where classical methods, like that of maximum likelihood, fail to give consistent estimators, it yields estimators which actually converge with probability one to the quantities being estimated; we call such estimators super-consistent. The problems treated in the present paper provide examples of this.

The basic ideas of the proofs of the super-consistency of these estimators are to be found in [2] and [4]. Application of the minimum distance method, unlike that of the method of maximum likelihood, is not mechanical, and, in the cases we have treated, always requires the development of special results.

The present paper presents results on problems not hitherto treated in the literature. It is intended to be largely self-contained, and its organization is as follows. Section 2 gives essential preliminaries. Section 3 contains a statement of some of the results already obtained elsewhere. In Section 4 are formulated three new problems in nonparametric stochastic difference equations. In Sections 5, 6, and 7 we exhibit minimum distance estimators for these problems. In Sections 5 and 6 we prove the super-consistency of the first two estimators.

In a few places the proofs are not given in all detail in the interest of brevity,

Received 6/24/53.

[1] This research was supported in part by the United States Air Force under Contract No. AF18(600)-685 monitored by the Office of Scientific Research.

Editor's Note: This paper was presented to the Chicago meeting of the Institute of Mathematical Statistics, December 29, 1952, and is published in the *Annals* by invitation of the Institute Committee on Special Invited Papers.

but sufficient detail is given to exhibit the fundamental ideas and spirit of the method. Places where the proofs below are not given in full detail are: Section 5, in the paragraph containing (5.10) and in the following paragraph; Section 6, for equations (6.3), (6.6), and (6.10), and in the paragraph following the one containing equation (6.10). At these points references to [2] and [4] are given where the reader will find similar theorems completely proved; a study of these proofs will enable him to reconstruct the missing points in all detail. The spirit of these results is discussed below when we discuss the basic ideas of the method. The proof of the result of Section 7 is omitted because it is easier than, and so much like, the proofs of Sections 5 and 6. Section 8 consists of concluding remarks.

The author is very grateful to Professor J. L. Doob for several helpful discussions while this paper was being written. Professors L. Hurwicz, T. Koopmans, and J. Marschak were very kind in answering the writer's questions about the literature and problems of stochastic difference equations.

The author wishes to take this opportunity to apologize for the inclusion, in the paper [2], of its Section 10. This section was by way of an incidental remark and had nothing to do with the minimum distance method. The idea of this section, as was kindly pointed out to the author by Professor W. Kruskal, was previously employed by Geary [5].

2. Essential preliminaries. Let $s_1, s_2, \cdots s_k$ be k numbers. By their empiric d.f. we mean a function, say $S(x)$, such that $kS(x)$ is equal to the number of these numbers $s_1, \cdots, s_k$ which are less than x. Let $(s_1, t_1), \cdots, (s_k, t_k)$, be k couples of numbers. By their empiric d.f. we mean a function, say $S(x, y)$, such that $kS(x, y)$ is the number of couples (s_i, t_i), $i = 1, \cdots, k$, such that $s_i < x$ and $t_i < y$. Similar definitions apply in higher dimensions.

Let (Z_1, Z_2) be a pair of chance variables. Their d.f. $G(x, y)$ is $P\{Z_1 < x$ and $Z_2 < y\}$, where $P\{\ \}$ denotes the probability of the relation in braces. Similar definitions apply in one and higher dimensions.

We stress here, as we have done in our previous papers, that our method does not depend upon any particular definition of distance, and is applicable with very many definitions. One of the problems requiring investigation is, in fact, to determine which definition will yield better results, and in what sense. Failing such knowledge, we will adopt the Fréchet distance which is mathematically convenient and not otherwise unreasonable.[2]

Let $S_1(x)$, $S_2(x)$ be a pair of d.f.'s. The distance $\delta(S_1, S_2)$ between them will be defined by

$$\delta(S_1, S_2) = \sup_x | S_1(x) - S_2(x) | .$$

[2] The notion of a metric space is due to Fréchet, so that in a large sense every distance is a Fréchet distance. We shall adopt the customary designation of the distance δ between two distribution functions as the Fréchet distance.

Similarly, the distance δ between the d.f.'s $S_3(x, y)$ and $S_4(x, y)$ will be defined by

$$\delta(S_2, S_4) = \sup_{x,y} | S_3(x, y) - S_4(x, y) | .$$

Similar definitions apply in higher dimensions. Let K be a class of d.f.'s. The distance $\delta(S_0, K)$ of the d.f. S_0 from the class K will be defined by

$$\delta(S_0, K) = \inf_{K' \epsilon K} \delta(S_0, K').$$

Let $Y_1, Y_2, \cdots$ be a sequence of independent, identically distributed chance variables with the common d.f. $G(x)$. Let $G_n^*(x)$ be the empiric d.f. of $Y_1, \cdots, Y_n$.

The theorem of Glivenko-Cantelli ([8], page 260) states:

$$P\left\{\lim_{n\to\infty} \delta[G(x), G_n^*(x)] = 0\right\} = 1. \tag{2.1}$$

Let $\{Y_{ij}\}, j = 1, \cdots, m_i, i = 1, 2, \cdots$ ad inf., be independently distributed chance variables. Let $G_i(x)$ be the common d.f. and $G_i^*(x)$ be the empiric d.f. of $Y_{i1}, \cdots, Y_{im_i}$. Define

$$G^n(x) = \frac{\sum_{i=1}^{n} m_i G_i(x)}{\sum_{i=1}^{n} m_i} \quad \text{and} \quad G^{n*}(x) = \frac{\sum_{i=1}^{n} m_i G_i^*(x)}{\sum_{i=1}^{n} m_i}.$$

An important tool in some applications of the minimum distance method is the following result (proved in [3]):

$$P\left\{\lim_{n\to\infty} \delta[G^n(x), G^{n*}(x)] = 0\right\} = 1. \tag{2.2}$$

(The approach to the limit in (2.2) is actually uniform in the G's; see Theorem 4.2 of [3]).

Let $\{Y_j^i\}, i = 1, \cdots, k; j = 1, 2, \cdots,$ ad inf., be a sequence of independent chance variables such that, for each i, $\{Y_j^i\}, j = 1, 2, \cdots$, ad inf., all have the same d.f. Let $q = (q_1, \cdots, q_k)$ be any k real parameters. Let $G(x \mid q)$ be the d.f. of $\sum_{i=1}^k q_i Y_1^i$, and $G_n(x \mid q)$ be the empiric d.f. of

$$\left\{\sum_{i=1}^{k} q_i Y_j^i\right\}, \qquad j = 1, \cdots, n.$$

Another important tool in the application of the minimum distance method is the following result (first proved in [4]):

$$P\left\{\lim_{n\to\infty} \sup_{q} \delta[G(x \mid q), G_n(x \mid q)] = 0\right\} = 1. \tag{2.3}$$

Actually this theorem is valid under much weaker hypotheses, and at the end of [4] there is given a prescription for proving this result under weaker conditions with essentially the same proof. In the new applications of the minimum distance method which we shall make later in this paper, we will actually make essential use of this theorem under several sets of weaker conditions. It should be noticed that this theorem does not merely say that the d.f. $G_n(x \mid q)$ converges to the d.f. $G(x \mid q)$ uniformly in q (actually the convergence of $G_n(x \mid q)$ to $G(x \mid q)$ is not only uniform in q but actually uniform in G (see the proof of Theorem 4.2 of [3])). The theorem actually says that the convergence is *simultaneous* for all q from $-\infty$ to $+\infty$, which is considerably more than uniformity.

3. Some previous results obtained by the minimum distance method. In this section we describe a few results already obtained, together with some heüristic considerations underlying them. We shall sometimes forego full generality in the interest of clarity of exposition.

Let $\{X_{ij}\}$, $i = 1, \cdots, n; j = 1, \cdots, m_i$ be independently distributed chance variables. (In [3] we discussed also the case where the chance variables are not independent). Let $F_i(x \mid \theta, \alpha_i)$ be the d.f. of $X_{i1}, \cdots, X_{im_i}$. The parameters θ and α_i upon which this d.f. depends are unknown; for simplicity we take them to be scalars although our results are equally valid for vectors. The parameter θ occurs in every group of X's ($X_{i1}, \cdots, X_{im_i}$ constitute the ith group) and was called by Neyman and Scott [6] "structural." The parameter α_i occurs only in the ith group and was called "incidental."

Let T be a (given) set within which θ is known to lie (of course T may be the whole line). Write

$$\bar{\alpha}_n = (\alpha_1, \alpha_2, \cdots, \alpha_n) \quad \text{and} \quad \bar{\alpha} = (\alpha_1, \alpha_2, \cdots).$$

Let A_n be the set within which $\bar{\alpha}_n$ is known to lie, and A the set within which $\bar{\alpha}$ is known to lie. Let $F_i^*(x)$ be the empiric distribution function of $X_{i1}, X_{i2}, \cdots, X_{im_i}$, and define

$$B^n(x) = \frac{\sum_{i=1}^n m_i F_i^*(x)}{\sum_{i=1}^n m_i}.$$

Let $\bar{\alpha}_n' = (\alpha_1', \cdots, \alpha_n')$, and define

$$C^n(x \mid \theta', \bar{\alpha}_n') = \frac{\sum_{i=1}^n m_i F_i(x \mid \theta', \alpha_i')}{\sum_{i=1}^n m_i}.$$

Let $\theta_n^*, \alpha_{1n}^*, \cdots, \alpha_{nn}^*$ be Borel-measurable functions of $X_{11}, \cdots, X_{nm_n}$ such that (writing $\alpha_n^* = (\alpha_{1n}^*, \cdots, \alpha_{nn}^*)$) $\theta_n^* \varepsilon T$, $\alpha_n^* \varepsilon A_n$, and

$$\delta[C^n(x \mid \theta_n^*, \alpha_n^*), B^n(x)] < \frac{1}{\text{n}} + \inf_{\theta' \varepsilon T, \bar{\alpha}_n' \varepsilon A_n} \delta[C^n(x \mid \theta', \bar{\alpha}_n'), B^n(x)]. \tag{3.1}$$

The estimator θ_n^* is a minimum distance estimator. Under a reasonable restriction it is proved in [3] that θ_n^* is a super-consistent estimator of θ. The basic ideas of this simple proof are as follows:

i) From (2.2) if follows that

$$P\left\{\lim_{n\to\infty} \delta[C^n(x \mid \theta, \bar{\alpha}_n)\ B^n(x)] = 0\right\} = 1. \tag{3.2}$$

Hence from the definition of θ^*_{n} it follows that a fortiori

$$P\left\{\lim_{n\to\infty} \delta[C^n(x \mid \theta^*_n, \alpha^*_n), B^n(x)] = 0\right\} = 1. \tag{3.3}$$

ii) If θ^*_n differs appreciably from θ then the distance

$$\delta[(C^n(x \mid \theta^*_n, \alpha^*_n), C^n(x \mid \theta, \bar{\alpha}_n)] \tag{3.4}$$

is appreciable. This is essentially the postulated restriction.

iii) Equations (3.2) and (3.3) imply that the distance (3.4) is almost always small for large n. Hence θ^*_n cannot differ appreciably from θ for large n.

We remind the reader that the above is only a heuristic outline of the proof, and also that the final result is a limiting property which holds with probability one.

Consider now the following problem: Let $\bar{\xi} = \xi_1, \xi_2, \cdots$ ad inf. be an infinite sequence of constants which are unknown to the statistician. Let α and β be parameters unknown to the statistician. Let (u_i, v_i), $i = 1, 2, \cdots$, ad inf., be a sequence of identically, independently, and jointly normally distributed pairs of chance variables, which the statistician cannot observe. The means of u_1 and v_1 are known to be zero; their covariance matrix is unknown. Let the observable chance variables be (x_i, y_i), $i = 1, \cdots, n$, where

$$x_i = \xi_i + u_i \quad \text{and} \quad y_i = \alpha + \beta\xi_i + v_i\,.$$

The problem is to give consistent estimators of α and β.

Let c_1 and c_2 be any real numbers and $A_n(x \mid c_1, c_2)$ be the empiric d.f. of $\{y_i - c_1 - c_2x_i\}$ for $i = 1, \cdots, n$. Let N^* be the class of all normal d.f.'s with mean zero. Define a_n and b_n as any Borel-measurable functions of the arguments $x_1, \cdots, x_n, y_1, \cdots, y_n$, such that

$$\delta[A_n(x \mid a_n, b_n), N^*] < \frac{1}{n} + \inf_{c_1, c_2} \delta[A_n(x \mid c_1, c_2), N^*].$$

It is proved in [3] under reasonable restrictions on the sequence $\bar{\xi}$ that a_n and b_n are super-consistent estimators of α and β, respectively. The basic ideas of this proof are as follows.

i) From (2.1) it follows that

$$P\left\{\lim_{n\to\infty} \delta[A_n(x \mid \alpha, \beta), N^*] = 0\right\} = 1. \tag{3.5}$$

Hence a fortiori we have

$$P\left\{\lim_{n\to\infty} \delta[A_n(x \mid a_n, b_n), N^*] = 0\right\} = 1. \tag{3.6}$$

ii) One proves that

$$P\left\{\lim_{n\to\infty} \delta\left[A_n(x \mid a_n, b_n), \frac{1}{n}\sum_{i=1}^{n} N(x \mid (\alpha - a_n) + (\beta - b_n)\xi_i, \sigma^2(b_n))\right] = 0\right\} = 1 \tag{3.7}$$

where $N(x \mid d_1, d_2)$ is the normal d.f. with mean d_1 and variance d_2,

$$\sigma^2(c) = \sigma_2^2 - 2c\rho\sigma_1\sigma_2 + c^2\sigma_1^2,$$

$$Eu^2 = \sigma_1^2, \qquad Ev^2 = \sigma_2^2, \qquad Euv = \rho\sigma_1\sigma_2 .$$

iii) One proves that, if $\mid \alpha - c_1 \mid + \mid \beta - c_2 \mid$ is appreciably different from zero, then the distance from N^* of

$$n^{-1}\sum_{i=1}^{n} N(x \mid (\alpha - c_1) + (\beta - c_2)\xi_i, \sigma^2(c_2)) \tag{3.8}$$

is appreciably different from zero.

From i, ii, and iii one concludes that a_n approaches α and b_n approaches β.

The postulated restrictions on ξ are such as to enable us to draw conclusions ii and iii. Meager restrictions suffice for this. In particular, if $\xi_1, \xi_2, \cdots$ are independent observations on a chance variable whose distribution is not normal (this is the case treated in [2]), these restrictions are satisfied with probability one. The conclusion of iii is proved by a compactness argument. In proving equation (3.7) one uses an argument similar to that used to prove (2.3) and first proves

$$P\left\{\lim_{n\to\infty} \sup_{c_1, c_2} \delta\left[A_n(x \mid c_1, c_2), n^{-1}\sum_{i=1}^{n} N(x \mid (\alpha - c_1) + (\beta - c_2)\xi_i, \sigma^2(c_2))\right] = 0 = \right\} 1. \tag{3.9}$$

From this (3.7) follows easily. The result (3.9) is much deeper and more difficult to prove than the result (2.2). One cannot obtain (3.7) directly from (2.2) because a_n and b_n are functions of $x_1, \cdots, x_n, y_1, \cdots, y_n$ and not constants. The relation (3.9) says not merely that (2.2) holds in this particular set-up uniformly in c_1, c_2, but actually *simultaneously* for all pairs c_1, c_2, which is considerably more than uniformly. This is essentially the relationship between (2.1) and (2.3). Mere uniformity is easy to prove but it is not what is needed.

In general, the proof of the super-consistency of a_n and b_n is much more difficult and elaborate than the corresponding proof for θ_n^*, chiefly in the need for proving (3.9). The operational reason seems to be the following: When estimating θ one has a definite empiric d.f. at his disposal, $(B_n(x))$, and compares it with the "true" d.f. and the nearest d.f. When estimating α and β one has to adjust the empiric d.f. $(A_n(x \mid c_1, c_2))$ until its distance from a sum of normal distributions which themselves depend upon the empiric d.f. is least. In the new problems treated below one obtains the estimator by varying a parameter until two empiric d.f.'s which depend upon it are closest together. One could then anticipate correctly that the proof of super-consistency will be more complicated as a result.

4. Statement of the new problems. In all that follows the indices i and n are to run through all the positive integers, and the index j is to run through all the integers, unless the contrary is explicitly stated. The sequence $\{u_j, v_j\}$ will always be a sequence of independent chance variables. All the u's are to have a common distribution, and all the v's are to have a common distribution. To avoid the trivial it will be assumed throughout this paper that neither u_1 nor v_1 is constant with probability one. The chance variables $\{u_j, v_j\}$ are statistically nonobservable variables. This means, mathematically speaking, that the estimators we shall construct will be functions of other (observable) variables, which will always be denoted by x_i.

PROBLEM A. Suppose

$$x_i = u_i + \alpha u_{i-1} \tag{4.1}$$

with α an unknown constant which may be any number less than one in absolute value. No assumption whatever will be made on the distribution of u_1. The problem is to estimate the parameter α; for all n we are to construct Borel measurable functions $a_n(x_1, \cdots, x_n)$ such that $a_n \to \alpha$ at least in probability. (Ours will converge to α with probability one).

PROBLEM B. Suppose β is an unknown constant which may be any number less than one in absolute value (other than zero), and

$$y_i = \beta y_{i-1} + u_i. \tag{4.2}$$

We wish this process to be stationary. It is easily seen that this implies that

$$y_i = \sum_{j=-\infty}^{i} \beta^{i-j} u_j. \tag{4.3}$$

Some assumption has now to be made so that the series in (4.3) will converge with probability one. Now let

$$x_i = y_i + v_i. \tag{4.4}$$

The problem is to construct estimators of the parameter β, that is, Borel measurable functions $b_n(x_1, \cdots, x_n)$ for all n such that $b_n \to \beta$; our estimators will converge with probability one.

PROBLEM C. Suppose γ is an unknown constant which may be any number less than one in absolute value, and

$$x_i = \gamma x_{i-1} + u_i. \tag{4.5}$$

The chance variable x_0 is chosen so as to make the process $\{x_i\}$ stationary. It is easily seen that this implies that

$$x_i = \sum_{j=-\infty}^{i} \gamma^{i-j} u_j. \tag{4.6}$$

We shall assume that 4.6 converges. The problem is to construct estimators of the parameter γ, that is, Borel-measurable functions $g_n(x_1, \cdots, x_n)$ for all n such that $g_n \rightarrow \gamma$; our estimators will converge with probability one.

Problems involving several simultaneous equations of higher order or problems of greater difficulty can of course also be treated by our minimum distance method. It seems to the author, however, that the attendant complications would obscure the essential points of the method. This explains our choice of problems.

5. Problem A. For convenience we will suppose in this section that the number of observations $\{x_i\}$ is odd and equal to $2n + 1$. Thus we will construct, for every positive n, a function $a_n(x_1, \cdots, x_{2n+1})$ of the arguments exhibited.

Let a be a real parameter which, in this section, will always be less than one in absolute value, and $A_n(x \mid a)$ be the empiric d.f. of $\{x_i - ax_{i-1}\}$ for $i = 2, \cdots, 2n + 1$. Define

$$B_n(x, y \mid a) = A_n(x \mid a) \cdot A_n(y \mid a). \tag{5.1}$$

Let $C_n(x, y \mid a)$ be the bivariate empiric d.f. of the pairs

$$\{(x_{2i+1} - ax_{2i}), (x_{2i} - ax_{2i-1})\} \qquad i = 1, \cdots, n.$$

Let a_n be any Borel-measurable function of $x_1, \cdots, x_{2n+1}$ such that $|a_n| < 1$ and

$$\delta[B_n(x, y \mid a_n), C_n(x, y \mid a_n)] < \frac{1}{n} + \inf_{|a|<1} \delta[B_n(x, y \mid a), C_n(x, y \mid a)]. \tag{5.2}$$

THEOREM 1. *We have*

$$P\{\lim_{n\rightarrow\infty} a_n = \alpha\} = 1 \tag{5.3}$$

that is, a_n is a super-consistent estimator of α.

The remainder of this section will be devoted to a proof of Theorem 1. Define

$$w_i(a) = x_i - ax_{i-1} = u_i + (\alpha - a)u_{i-1} - a\alpha u_{i-2}. \tag{5.4}$$

Hence

$$w_i(\alpha) = u_i - \alpha^2 u_{i-2}. \tag{5.5}$$

We see that $w_i(a)$ and $w_{i'}(a)$ are independently distributed whenever $|i - i'| \geqq 3$. Also, for any i, $w_i(\alpha)$ and $w_{(i+1)}(\alpha)$ are independently distributed.

Let $H(x \mid a)$ be the d.f. of $w_i(a)$. Let $A_{in}(x \mid a)$, $i = 1, 2, 3$, be the empiric d.f. of all $w_j(a)$ such that $2 \leqq j \leqq 2n + 1$, and $j \equiv i \pmod 3$. Thus each $A_{in}(x \mid a)$ is the empiric d.f. of independently distributed chance variables $(w_j(a))$, each of which is the same linear combination of independently (within each sequence) distributed u_i's. From our generalization [4] of the theorem of Glivenko-Cantelli we obtain that

$$P\{\lim_{n\rightarrow\infty} \sup_{|a|<1} \delta[A_{in}(x \mid a), H(x \mid a)] = 0\} = 1 \qquad \text{for } i = 1, 2, 3, \tag{5.6}$$

Hence

$$P\left\{\lim_{n\to\infty}\sup_{|a|<1}\delta[A_n(x\mid a),\,H(x\mid a)]=0\right\}=1. \tag{5.7}$$

Let

$$D(x,y\mid a)=H(x\mid a)\cdot H(y\mid a). \tag{5.8}$$

From (5.7) we obtain

$$P\left\{\lim_{n\to\infty}\sup_{|a|<1}\delta[B_n(x,y\mid a),\,D(x,y\mid a)]=0\right\}=1. \tag{5.9}$$

Let $E(x, y \mid a)$ be the d.f. of $(w_3(a), w_2(a))$. Then the infimum of

$$\delta[E(x,y\mid a),\,D(x,y\mid a)]$$

in the domain $\{|a| < 1, |a - \alpha| \geqq d > 0\}$ is, say, $l(d) > 0$. This is proved by a compactness argument based on the following two facts:

i) If $l(d) = 0$ then, for some number a_0 with $|a_0 - \alpha| \geqq d$ and $|a_0| \leqq 1$,

$$E(x,y\mid a_0)\equiv D(x,y\mid a_0), \tag{5.10}$$

ii) But this cannot hold because, when $(\alpha - a)(1 - |a\alpha|) \neq 0$ (as is surely the case for $a = a_0$), $w_3(a)$ and $w_2(a)$ are not independently distributed, as (5.8) would then imply. To show the latter we employ the following argument. Let $\varphi(t)$ be the logarithm of the characteristic function of u_1. This is well defined in a neighborhood of the origin, which is the only place where we will require $\varphi(t)$; we use that branch of the function $\varphi(t)$ for which $\varphi(0) = 0$. The independence of $w_3(a)$ and $w_2(a)$ would imply that, in a neighborhood of the origin,

$$\varphi([\alpha-a]s+t)+\varphi(-a\alpha s+[\alpha-a]t)$$
$$=\varphi([\alpha-a]s)+\varphi(t)+\varphi(-a\alpha s)+\varphi([\alpha-a]t).$$

If now

$$\varphi([\alpha-a]s+t)=\varphi([\alpha-a]s)+\varphi(t)$$

for all s and t in a neighborhood of the origin, then $\varphi(s) = c_0 s$ in a neighborhood of the origin, where c_0 is a constant. Since $\exp\{\varphi(s)\}$ is a characteristic function at least for small $|s|$, it follows that c_0 is purely imaginary and hence that the characteristic function of u_1 is $\exp\{c_0 s\}$ for all s. This violates the assumption that u_1 is not constant with probability one and proves the desired result. We now employ the following argument due to J. L. Doob. Define (always only in a neighborhood of the origin) the function

$$\psi(s,t)=\varphi(s+t)-\varphi(s)-\varphi(t).$$

Then $\psi(s, t)$ is continuous and $\psi(s, t) = \psi(t, s)$. Also $\psi(0, t) = 0$ and

$$\psi([\alpha-a]s,t)+\psi(-a\alpha s,[\alpha-a]t)\equiv 0.$$

Hence

$$\begin{aligned}\psi(s, t) &= -\psi(-a\alpha[\alpha - a]^{-1}s, [\alpha - a]t)\\ &= -\psi([\alpha - a]t, -a\alpha[\alpha - a]^{-1}s)\\ &= \psi(-a\alpha t, -a\alpha s) = \psi(-a\alpha s, -a\alpha t).\end{aligned}$$

Now always $| a\alpha | < 1$ since $| \alpha | < 1$. For every positive integer n we have

$$\psi(s, t) = \psi((-a\alpha)^n s, (-a\alpha)^n t).$$

Hence $\psi(s, t) \equiv 0$ in a neighborhood of the origin, so that the proof is complete.

Essentially as in [4], making use of the fact that each w_i is a linear combination of independent u's, one can prove that

$$P\left\{\limsup_{n\to\infty}\ \sup_{|a|<1} \delta[C_n(x, y \mid a), E(x, y \mid a)] = 0\right\} = 1. \tag{5.11}$$

The facts cited in the last two paragraphs are basic to our proof of convergence. Theorems corresponding to them are proved in [2] and [4] and cited below for our other problems. The method of proof for new problems will consist in part of choosing suitable d.f.'s for which one can assert similar theorems. The proofs of the present theorems require considerable detail, but can be constructed by the reader who understands the ideas of the proofs in [2] and [4]. They are omitted here and in subsequent sections of this paper because their detailed exposition would make this paper inordinately long for both reader and writer.

Suppose now that the theorem is not true. Then there exist positive d_1 and d_2 such that

$$P\left\{\limsup_{n\to\infty} | a_n - \alpha | > d_1\right\} > d_2 . \tag{5.12}$$

Hence

$$P\left\{\limsup_{n\to\infty} \delta[E(x, y \mid a_n), D(x, y \mid a_n)] \geqq l(d_1)\right\} > d_2 . \tag{5.13}$$

From (5.11) we obtain

$$P\left\{\lim_{n\to\infty} \delta[C_n(x, y \mid a_n), E(x, y \mid a_n)] = 0\right\} = 1. \tag{5.14}$$

From (5.13) and (5.14) we obtain

$$P\left\{\limsup_{n\to\infty} \delta[C_n(x, y \mid a_n), D(x, y \mid a_n)] \geqq l(d_1)\right\} > d_2 . \tag{5.15}$$

From (5.9) and (5.15) we obtain

$$P\left\{\limsup_{n\to\infty} \delta[B_n(x, y \mid a_n), C_n(x, y \mid a_n)] \geqq l(d_1)\right\} > d_2 . \tag{5.16}$$

Since $w_i(\alpha)$ and $w_{i+1}(\alpha)$ are independently distributed we have

$$E(x, y \mid \alpha) \equiv D(x, y \mid \alpha). \tag{5.17}$$

From (5.9) therefore we obtain

$$(5.18)\qquad P\left\{\lim_{n\to\infty} \delta[B_n(x, y \mid \alpha), E(x, y \mid \alpha)] = 0\right\} = 1.$$

From 5.11) and (5.18) we obtain

$$(5.19)\qquad P\left\{\lim_{n\to\infty} \delta[B_n(x, y \mid \alpha), C_n(x, y \mid \alpha)] = 0\right\} = 1.$$

From (5.2) and (5.19) we obtain

$$(5.20)\qquad P\left\{\lim_{n\to\infty} \delta[B_n(x, y \mid a_n), C_n(x, y \mid a_n)] = 0\right\} = 1.$$

The contradiction between (5.16) and (5.20) proves Theorem 1.

6. Problem B. For convenience we will assume in this section that the number of observations $\{x_i\}$ is $4n + 1$. Thus we will construct, for every n, a function $b_n(x_1, \cdots, x_{4n+1})$ of the arguments exhibited.

Let b be a real parameter which throughout this section will be assumed to be less than one in absolute value. Let

$$m_i(b) = x_i - bx_{i-1} = u_i + (v_i - bv_{i-1}) + (\beta - b)y_{i-1}.$$

If $| i - i' | \geqq 2$ then $m_i(\beta)$ and $m_{i'}(\beta)$ are independently distributed. If $b \neq \beta$ and $| i - i' | \geqq 2$ then $m_i(b)$ and $m_{i'}(b)$ are not independently distributed.

Let $A_n(x \mid b)$ be the empiric d.f. of $m_2(b), \cdots, m_{4n+1}(b)$. Define

$$B_n(x, y \mid b) = A_n(x \mid b)A_n(y \mid b).$$

Let $C_n(x, y \mid b)$ be the bivariate empiric d.f. of the $2n$ pairs

$$(m_2(b), m_4(b)); \qquad (m_3(b), m_5(b)); \qquad (m_6(b), m_8(b));$$
$$(m_7(b), m_9(b)); \cdots; \qquad (m_{4n-1}(b), m_{4n+1}(b)).$$

Let b_n be any Borel measurable function of $x_1, \cdots, x_{4n+1}$ such that $| b_n | < 1$ and

$$(6.1)\qquad \delta[B_n(x, y \mid b_n), C_n(x, y \mid b_n)] < \frac{1}{n} + \inf_b \delta[B_n(x, y \mid b), C_n(x, y \mid b)].$$

We shall now sketch the proof of

THEOREM 2. *We have*

$$(6.2)\qquad P\{\lim_{n\to\infty} b_n = \beta\} = 1.$$

For any b, the sequence $\{m_i(b)\}$ for $i = 2, 3, \cdots$, ad inf., is a sequence of stationary chance variables. Moreover, each $m_i(b)$ is a (stationary) linear combination of u's and v's which are all independently distributed. Hence the stochastic process $\{m_i(b)\}$ is metrically transitive, for any b. Making use of the ergodic theorem we obtain without difficulty that the conclusion of the Glivenko-Cantelli theorem ([8], page 260) holds for the sequence $\{m_i(b)\}$, whatever be b.

Let $H(x \mid b)$ be the d.f. of $m_i(b)$. Using the methods of [2] and [4] one can prove that

$$P\{\lim_{n\to\infty} \sup_b \delta[A_n(x \mid b), H(x \mid b)] = 0\} = 1. \tag{6.3}$$

Let

$$D(x, y \mid b) = H(x \mid b)\cdot H(y \mid b). \tag{6.4}$$

From (6.3) we obtain

$$P\{\lim_{n\to\infty} \sup_b \delta[B_n(x, y \mid b), D(x, y \mid b)] = 0\} = 1. \tag{6.5}$$

Let $C_{1n}(x, y \mid b)$ be the bivariate empiric d.f. of the pairs

$$(m_2(b), m_4(b)); \quad (m_6(b), m_8(b)); \quad (m_{10}(b), m_{12}(b)); \cdots; \quad (m_{4n-2}(b), m_{4n}(b)).$$

and $C_{2n}(x, y \mid b)$ be the bivariate empiric d.f. of the pairs

$$(m_3(b), m_5(b)); \quad (m_7(b), m_9(b)), \cdots; \quad (m_{4n-1}(b), m_{4n+1}(b)).$$

When $|i - i'| \geqq 2$, $m_i(\beta)$ and $m_{i'}(\beta)$ are independently distributed. Hence from the extension to the present case of the bivariate Glivenko-Cantelli theorem we obtain

$$P\{\lim_{n\to\infty} \delta[C_{in}(x, y \mid \beta), D(x, y \mid \beta)] = 0\} = 1, i = 1, 2. \tag{6.6}$$

Hence

$$P\{\lim_{n\to\infty} \delta[C_n(x, y \mid \beta), D(x, y \mid \beta)] = 0\} = 1. \tag{6.7}$$

From (6.5) and (6.7) we obtain

$$P\{\lim_{n\to\infty} \delta[B_n(x, y \mid \beta), C_n(x, y \mid \beta)] = 0\} = 1. \tag{6.8}$$

From (6.1) and (6.8) we obtain

$$P\{\lim_{n\to\infty} \delta[B_n(x, y \mid b_n), C_n(x, y) \mid b_n)] = 0\} = 1. \tag{6.9}$$

Using the methods of [2] and [4] it can be proved,[3] although considerable detail is required, that, for $i = 1, 2$,

$$P\{\lim_{n\to\infty} \sup_b \delta[C_{in}(x, y \mid b), E(x, y \mid b)] = 0\} = 1, \tag{6.10}$$

where $E(x, y \mid b)$ is the d.f. of the pair $(m_2(b), m_4(b))$. From (6.10) we obtain

$$P\{\lim_{n\to\infty} \sup_b \delta[C_n(x, y \mid b), E(x, y \mid b)] = 0\} = 1. \tag{6.11}$$

[3] This illustrates the fact that the result of [4] does not require for its validity the independence of the chance variables. It is actually valid under much weaker conditions and obviously can be extended to multivariate distributions.

By a compactness argument similar to that of Lemma 2 of [2] one can prove that the infimum of $\delta[E(x,y \mid b), D(x,y \mid b)]$ in the domain $\{\mid b \mid < 1, \mid b - \beta \mid \geqq d > 0\}$ is, say, $l_1(d) > 0$. (The reader will have noticed that $E(x, y \mid \beta) \equiv D(x, y \mid \beta)$, and, for $b \neq \beta$, that $\delta[E(x, y \mid b), D(x, y \mid b)] > 0$.)

Suppose now that Theorem 2 is not true and there exist positive d_1 and d_2 such that

$$P\ \{ \limsup_{n\to\infty} \mid b_n - \beta \mid > d_1 \} > d_2. \tag{6.12}$$

Hence

$$P\ \{ \limsup_{n\to\infty} \delta[E(x, y \mid b_n), D(x, y \mid b_n)] \geqq l_1(d_1) \} > d_2. \tag{6.13}$$

From (6.11) and (6.13) we obtain

$$P\ \{ \limsup_{n\to\infty} \delta[C_n(x, y \mid b_n), D(x, y \mid b_n)] \geqq l_1(d_1) \} > d_2. \tag{6.14}$$

Together with (6.5) this yields

$$P\ \{ \limsup_{n\to\infty} \delta[B_n(x, y \mid b_n), C_n(x, y \mid b_n)] \geqq l_1(d_1) \} > d_2. \tag{6.15}$$

The contradiction between (6.9) and (6.15) proves Theorem 2.

7. Problem C. For convenience we will assume in this section that the number of observations $\{x_i\}$ is odd and $2n + 1$, say. Thus we will construct, for every n, a function $g_n(x_1, \cdots, x_{2n+1})$ of the arguments exhibited.

Let g be a real parameter which throughout this section will be assumed to be less than one in absolute value. Let

$$q_i(g) = x_i - gx_{i-1} = u_i + (\gamma - g)x_{i-1}. \tag{7.1}$$

The chance variables $\{q_i(\gamma)\}$ are all independent of each other. If $g \neq \gamma$ and $i \neq i'$ then $q_i(g)$ and $q_{i'}(g)$ are not independently distributed.

Let $A_n(x \mid g)$ be the empiric d.f. of $q_2(g), q_3(g), \cdots, q_{2n+1}(g)$. Define

$$B_n(x, y \mid g) = A_n(x \mid g) \cdot A_n(y \mid g).$$

Let $C_n(x, y \mid g)$ be the bivariate empiric d.f. of the pairs

$$(q_2(g), q_3(g)); \qquad (q_4(g), q_5(g)); \cdots; \qquad (q_{2n}(g), q_{2n+1}(g)).$$

Let g_n be any Borel-measurable function of $x_1, \cdots, x_{2n+1}$ such that $\mid g_n \mid < 1$ and

$$\delta[B_n(x, y \mid g_n), C_n(x, y \mid g_n)] < \frac{1}{n} + \inf_g \delta[B_n(x, y \mid b), C_n(x, y \mid b)]. \tag{7.2}$$

Then, in a manner similar to that of preceding sections, one can prove that

$$P\ \{ \lim_{n\to\infty} g_n = \gamma \} = 1.$$

8. Conclusion. What the "practical" value of the minimum distance method is is very unclear at present to the writer. For example, the method enables one (Section 3 or [2] and [3]) to fit a straight line when both variables are subject to normal[4] errors, under an assumption (on ξ) so weak that the very pretty result of Reiersøl [9] is an immediate consequence. (Reiersøl's theorem states that, if the $\xi_1, \xi_2, \cdots$ of Section 3 are independent chance variables with a common distribution function which is not normal, then α and β are identified). However, if one assumes that any cumulant of order not less than three of the common distribution is not zero—an assumption to which many practical people would not object—one can, using Geary's method ([5] or [2], Section 10) expeditiously obtain consistent estimators of α and β. It might therefore be argued that the difficulty of the problem is due solely to insistence on mathematical generality and aesthetics, and disappears when one is willing to make practical assumptions. The same argument could be made about the problems described in Section 4 of this paper; if one assumes second moments to exist one can, without any difficulty, obtain consistent estimators.

It seems to the writer, however, that the minimum distance method is of interest precisely because it enables one to solve a class of problems which cannot be solved by classical methods, and to do this in a manner which seems very reasonable and suggestive. The problems need not be solely problems of estimation but may also be problems of testing hypotheses. Thus (see [2], page 149) suppose one wishes to test the hypothesis that the common distribution function of the independent chance variables $z_1, \cdots, z_n$ is normal. One could base this test on $\delta(Z_n, N^{**})$, where $Z_n(x)$ is the empiric distribution function of $z_1, \cdots, z_n$, and N^{**} is the class of all normal distribution functions. Also there is no doubt that the minimum distance method is useful in the solution of many identification problems (for a discussion of identification problems see Koopmans [7]). Reiersøl's theorem and other problems of [3] and the present paper are cases in point. It is the author's opinion that the minimum distance method will also be useful in the treatment of many nonparametric problems.

An important general problem is to find a method of full generality which will yield efficient estimators of structural parameters in the case where each new set of observations depends also upon another incidental parameter. The solution of this problem is at present unknown. Neyman and Scott [6] have shown that the method of maximum likelihood does not always yield efficient or even consistent estimators. The minimum distance method as employed in [3] (briefly described in Section 3 of the present paper) yields consistent estimators in rather wide generality; its efficiency remains to be determined.

If an efficient estimator does not exist the problem would seem to be to characterize the complete class of estimators. One should not a priori preclude the possibility of employing some reasonable measure of efficiency other than the

[4] Actually, as pointed out in [2], the errors need not be normal, nor need the linear relation be in two dimensions only.

usual one. If most or many consistent estimators are not normally distributed this may be advisable.

Among statistical methods which employ the idea of distance is the one for which Kolmogoroff and Smirnoff obtained many asymptotic distributions and for which Wald and the present writer obtained small sample results (for a description and references see, for example, Birnbaum [10]). Suppose, for example, that one wishes to test the simple hypothesis that the distribution function of n independent, identically distributed chance variables is a given distribution function $F(x)$. The Kolmogoroff-Smirnoff test is based on the Fréchet distance between $F(x)$ and the empiric distribution function of the chance variables. There is no minimization of distance in the Kolmogoroff-Smirnoff test. In the application of the minimum distance method one always minimizes the distance between two distribution functions, or between a distribution function and a class of distribution functions, or between two classes of distribution functions.

REFERENCES

[1] J. Wolfowitz, "The method of maximum likelihood and the Wald theory of decision functions," *Proc. Roy. Dutch Acad. Sci.*, Vol. 56 (1953), pp. 114–119.

[2] J. Wolfowitz, "Consistent estimators of the parameters of a linear structural relation," *Skand. Aktuarietids.*, Vol. 35 (1952), pp. 132–151.

[3] J. Wolfowitz, "Estimation by the minimum distance method," *Ann Inst. Stat. Math.*, Tokyo, Vol. 5 (1953), pp. 9–23.

[4] J. Wolfowitz, "Generalization of the theorem of Glivenko-Cantelli," *Ann. Math. Stat.* Vol. 25 (1954), pp. 131–138.

[5] R. C. Geary, "Inherent relations between random variables," *Proc. Roy. Irish Acad.*, Vol. 47, A. 6 (1942), p. 195.

[6] J. Neyman and E. L. Scott, "Consistent estimators based on partially consistent observations," *Econometrica*, Vol. 16 (1948), pp. 1–32.

[7] T. C. Koopmans, editor. *Statistical Inference in Dynamic Economic Models*, John Wiley and Sons, 1950.

[8] M. Fréchet, *Recherches Théoriques Modernes sur le Calcul des Probabilités*, Vol. 1, Gauthier-Villars, Paris, 1937.

[9] O. Reiersøl, "Identifiability of a linear relation between variables which are subject to error," *Econometrica*, Vol. 18 (1950), pp. 375–389.

[10] Z. W. Birnbaum, "Distribution-free tests of fit for continuous distribution functions," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 1–8.

Note added in proof. The author has recently succeeded in applying the minimum distance method, in a manner different from that of the present paper, to a considerably larger class of problems. Linearity or other such restrictions are not needed, application is fairly routine, the proofs are much simpler, and the result of [4] is not used. Identified distribution functions can also be estimated. A brief description of these results will appear approximately concurrently with the present paper in the *Proceedings of the National Academy of Sciences.*