CONSISTENCY OF THE MAXIMUM LIKELIHOOD ESTIMATOR IN THE PRESENCE OF INFINITELY MANY INCIDENTAL PARAMETERS

By J. Kiefer[1] and J. Wolfowitz[2]

Cornell University

Summary. It is shown that, under usual regularity conditions, the maximum likelihood estimator of a structural parameter is strongly consistent, when the (infinitely many) incidental parameters are independently distributed chance variables with a common unknown distribution function. The latter is also consistently estimated although it is not assumed to belong to a parametric class. Application is made to several problems, in particular to the problem of estimating a straight line with both variables subject to error, which thus after all has a maximum likelihood solution.

1. Introduction. Let $\{X_{ij}\}$, $i = 1, \cdots, n$, $j = 1, \cdots, k$, be chance variables such that the frequency function of $X_{i1}, \cdots, X_{ik}$ is $f(x \mid \theta, \alpha_i)$ when θ and α_i are given, and thus depends upon the unknown (to the statistician) parameters θ and α_i. The parameter θ, upon which all the distributions depend, is called "structural"; the parameters $\{\alpha_i\}$ are called "incidental". Throughout this paper we shall assume that the X_{ij} are independently distributed when $\theta, \alpha_1, \cdots, \alpha_n$, are given, and shall consider the problem of consistently estimating θ (as $n \to \infty$). The chance variables $\{X_{ij}\}$ and the parameters θ and $\{\alpha_i\}$ may be vectors. However, for simplicity of exposition we shall throughout this paper, except in Example 2, assume that they are scalars. Obvious changes will suffice to treat the vector case.

Very many interesting problems are subsumed under the above formulation. Among these is the following:

$$f(x \mid \theta, \alpha_i) = (2\pi\theta)^{-k/2} \exp\left\{\frac{-\sum_j (x_{ij} - \alpha_i)^2}{2\theta}\right\}. \tag{1.1}$$

Suppose now that the $\{\alpha_i\}$ are considered as unknown constants and we form in the usual manner the likelihood function

$$(2\pi\theta)^{-kn/2} \exp\left\{-\frac{1}{2\theta}\sum_{i,j} (X_{ij} - \alpha_i)^2\right\} \tag{1.2}$$

corresponding to (1.1). Then the maximum likelihood (m.l.) estimator of θ is

$$\frac{\sum_{i,j} (X_{ij} - \bar{X}_i)^2}{kn} \tag{1.3}$$

Received September 28, 1955.

[1] Research sponsored by the Office of Naval Research.

[2] The research of this author was supported in part by the United States Air Force under Contract AF 18(600)-685, monitored by the Office of Scientific Research.

with $\bar{X}_i = k^{-1}\sum_j X_{ij}$, and is obviously not consistent. This example is due to Neyman and Scott [1], who used it to prove that the m.l. estimator[3] need not be consistent when there are infinitely many incidental parameters (constants). The latter authors, to whom the names "structural" and "incidental" are due, seem to have been the first to formulate the general problem. Special forms of the problem, like Example 2 below, had been studied for a long time (e.g., Wald [2] and the literature cited there).

The general fact that, when the $\{\alpha_i\}$ are unknown constants, the m.l. estimator of θ need not be consistent, is certainly basically connected with the fact that, since there are only a constant number of observations which involve a particular α_i, it is in general impossible to estimate the $\{\alpha_i\}$ consistently. Now there are many meaningful and practical statistical problems where the $\{\alpha_i\}$ are not arbitrary constants but independently and identically distributed chance variables with distribution function (df) G_0 (unknown to the statistician). The question then arises whether the m.l. method, which does not always yield a consistent estimator when there are infinitely many incidental constants, and does yield consistent estimators in the classical parametric case where there are no incidental parameters, will give a consistent estimator in this case, where the $\{\alpha_i\}$ are independent chance variables with the common df G_0. This note is devoted to this question.

The answer is affirmative. Not only is the m.l. estimator of θ strongly consistent (i.e., converges to θ with probability one) under reasonable regularity conditions, but also the m.l. estimator of G_0 converges to G_0 at every point of continuity of the latter, with probability one (w.p.1). This is the more striking when one recalls that G_0 does not belong to a parametric class, i.e., a set of df's indexed by a finite number of parameters. (If G_0 were a member of such a given class, the problem would fall completely in the domain of classical maximum likelihood.) The interest of the present authors was originally in estimating θ. That G can also be estimated by the m.l. method is a felicitous by-product of our investigation. A heuristic explanation of the present result may be this: A sequence of chance variables is more "regular" than an arbitrary sequence of numbers. In the present procedure one does not attempt to determine the particular values of the chance variables $\{\alpha_i\}$, but only their distribution function; thus, we seek the m.l. estimator of the "parameter" $\gamma = (\theta, G)$ based on a sequence of independent random variables whose common distribution function is indexed by γ.

In sections 3, 4, and 5, the results are applied to three problems which seem to be of interest per se. Among these is the problem of fitting a straight line with both variables subject to normal error. This problem has a very long history and has been the subject of many investigations (see, for example [2], [7], [4], and the literature cited there); it seems interesting that it can, after all, be treated by the m.l. method. The verification of the regularity assumptions or the formulation of not too onerous conditions for them to be verified is sometimes not entirely ob-

[3] Throughout this paper, for the sake of brevity, we use the term "estimator" to mean "sequence of estimators for $n = 1, 2, \cdots$."

vious, and the verification of these assumptions (in the form used in Section 2) constitutes the main difficulty of the paper. As is explained in detail below, the fact that these assumptions imply the general consistency result of Section 2 follows from a modification of the proof of [5]. Professor Herbert Robbins has kindly called our attention to his abstract in *Ann. Math. Stat.*, vol. 21 (1950), p. 314, Abstract 35, which states that the m.l. estimator of G is consistent. Since nothing further has appeared on this subject, the intended restrictions under which the statement is true, and the intended method of proof, are unknown to the present authors. This seems to be the second instance in the literature where the m.l. estimator has been used to estimate an entire df which is not assumed to belong to a class depending only on a finite number of real parameters. The first instance of the employment of such an estimator is the classical estimation of a df by its empiric df (shown to be asymptotically optimal in [3]), which is its m.l. estimator (see the paragraph preceding the lemma in Section 2). The only other instance of the estimation of a df in the nonparametric case seems to be that of the estimation of identifiable df's in stochastic structures such as those of the present paper by means of the minimum distance method [4].[4] (The latter requires regularity conditions weaker than those of the present paper. Compare, for example, [4] with Example 2 below; see also Example 3a.)

In connection with these examples, and also in Section 6, we give some examples which illustrate the fact that the classical m.l. estimator may not be consistent, even in parametric examples which lack the pathological discontinuity sometimes present in hitherto published examples.

Section 6 also contains the statement of a simple device which can be used in the classical parametric case as well as in the case studied in this paper, to prove consistency of the m.l. estimator in some cases where the assumptions used in published proofs of consistency are not satisfied.

The proof in Section 2 is a modification of Wald's [5], and its fundamental ideas are to be found in [5]; for this reason some of its details will be omitted. Wald states in his paper that his method applies more generally when his Assumption 9 is fulfilled. However, this assumption is not fulfilled in our problem *ab initio* and some technical modifications have to be made. One obstacle to extending Wald's proof to our problem is in establishing an analogue of (16) in [5]; one "neighborhood of infinity" does not always seem to suffice. Also some changes in the assumptions are made necessary by the nature of our problem. The results of the present paper can be extended in the usual manner to abstract spaces, but we forego this. It should also be remarked that in [6] Wald studied the present problem of estimating a structural parameter.

The attitude towards the $\{\alpha_i\}$, i.e., whether they are to be regarded as unknown constants or identically and independently distributed chance variables or something else, seems to vary with the author and sometimes even within the

[4] A paper entitled "The minimum distance method," which gives the details and proofs of the results announced in [4], is scheduled for publication in a forthcoming issue of these Annals.

publications of the same author. For example, Wald [2], in his treatment of the problem of fitting a straight line mentioned above, considers the $\{\alpha_i\}$ as unknown constants; and Neyman and Scott, in their general formulation of the problem given in [1] and described at the beginning of the present section, also consider the $\{\alpha_i\}$ as unknown constants. On the other hand, Neyman in his treatment [7] of the straight line problem treats the α_i as independently and identically distributed chance variables. Also Neyman and Scott [8] criticize Wald's solution [2] on the ground that the conditions he postulates on the sequence of constants $\{\alpha_i\}$ are such that they are unlikely to be satisfied when the $\{\alpha_i\}$ are independently and identically distributed chance variables. Our own point of view and perhaps also that of the other writers cited, is that one need not insist on any one formulation to the exclusion of all others. There are certainly reasonable statistical problems where the $\{\alpha_i\}$ may be looked upon as independently and identically distributed chance variables, and consequently the problem of the present paper is statistically meaningful and interesting. This is also the attitude implicit in [4] and [9].

2. Proof of consistency. As we have stated earlier, the essential idea of the proof comes from [5]. A compactification device has to be employed because the space Γ defined below may not be compact.

We postulate that the following assumptions are fulfilled (see also the paragraph preceding the lemma at the end of this section):

ASSUMPTION 1: $f(x \mid \theta, \alpha)$ is a density with respect to a σ-finite measure μ on a Euclidean space of which x is the generic point. (This is also Wald's Assumption 1.)

Let Ω be the space of possible values of θ, and let A be the space of values which α_i can take. (Both Ω and A are measurable subsets of Euclidean spaces, f is jointly measurable in x and α for each θ, and we hereafter denote by $\theta_i^{(s)}$ $(1 \leqq s \leqq r)$ the components of a point θ_i in Ω and by $\mid \alpha \mid$ the Euclidean distance from the origin of a point $\alpha \in A$; τ will denote Lebesgue measure on A.) Let $\Gamma = \{G\}$ be a given space of (cumulative) distributions of α_i. Let θ_0, G_0 be, respectively, the "true" value of the parameter θ and the "true" distribution of α_i. It is assumed that $\theta_0 \in \Omega$ and $G_0 \in \Gamma$. Let $\gamma = (\theta, G)$ be the generic point in $\Omega \times \Gamma$. We define

$$f(x \mid \gamma) = \int_A f(x \mid \theta, z)\, dG(z) \tag{2.1}$$

and $\gamma_0 = (\theta_0, G_0)$. In the space $\Omega \times \Gamma$ we define the metric

$$\begin{aligned} \delta(\gamma_1, \gamma_2) &= \delta([\theta_1, G_1], [\theta_2, G_2]) \\ &= \sum_{s=1}^{r} \mid \arctan \theta_1^{(s)} - \arctan \theta_2^{(s)} \mid \\ &\quad + \int_A \mid G_1(z) - G_2(z) \mid e^{-\mid z \mid}\, d\tau(z). \end{aligned} \tag{2.2}$$

Let $\bar{\Omega} \times \bar{\Gamma}$ be the completed space of $\Omega \times \Gamma$ (the space together with the limits of its Cauchy sequences in the sense of the metric (2.2)). Then $\bar{\Omega} \times \bar{\Gamma}$ is compact.

ASSUMPTION 2 (Continuity Assumption): It is possible to extend the definition of $f(x \mid \gamma)$ so that the range of γ will be $\bar{\Omega} \times \bar{\Gamma}$ and so that, for any $\{\gamma_i\}$ and γ^* in $\bar{\Omega} \times \bar{\Gamma}$, $\gamma_i \to \gamma^*$ implies

$$f(x \mid \gamma_i) \to f(x \mid \gamma^*), \tag{2.3}$$

except perhaps on a set of x whose probability is 0 according to the probability density $f(x \mid \gamma_0)$. (The exceptional x-set may depend on γ^*; $f(x \mid \gamma^*)$ need not be a probability density function.) (This assumption corresponds to Wald's continuity Assumptions 3 and 5.)

ASSUMPTION 3: For any γ in $\bar{\Omega} \times \bar{\Gamma}$ and any $\rho > 0$, $w(x \mid \gamma, \rho)$ is a measurable function of x, where

$$w(x \mid \gamma, \rho) = \sup f(x \mid \gamma'),$$

the supremum being taken over all γ' in $\bar{\Omega} \times \bar{\Gamma}$ for which $\delta(\gamma, \gamma') < \rho$. (This assumption is made for the reasons given by Wald. See his remarks following Assumption 8 in [5].)

ASSUMPTION 4 (Identifiability Assumption): If γ_1 in $\bar{\Omega} \times \bar{\Gamma}$ is different from γ_0, then, for at least one y,

$$\int_{-\infty}^{y} f(x \mid \gamma_1)\, d\mu \neq \int_{-\infty}^{y} f(x \mid \gamma_0)\, d\mu, \tag{2.4}$$

the integral being over those x all of whose components are $\leqq$ the corresponding components of y. (This is the same as Wald's Assumption 4.)

Let X be a chance variable with density $f(x \mid \gamma_0)$. *The operator E will always denote expectation under γ_0*; γ_0 will always, of course, be a member of $\Omega \times \Gamma$.

ASSUMPTION 5 (Integrability Assumption): For any γ in $\bar{\Omega} \times \bar{\Gamma}$ we have, as $\rho \downarrow 0$,

$$\lim E\left[\log \frac{w(X \mid \gamma, \rho)}{f(X \mid \gamma_0)}\right]^+ < \infty. \tag{2.5}$$

(This assumption is implied by assumptions corresponding to Wald's Assumptions 2 and 6.)

For any γ in $\bar{\Omega} \times \bar{\Gamma}$ other than γ_0, define $v = \log f(X, \gamma) - \log f(X, \gamma_0)$. We begin the proof of consistency by showing that

$$Ev < 0. \tag{2.6}$$

First, if γ is in $\Omega \times \Gamma$, $Ee^v \leqq 1$. Hence from (2.3) and Fatou's lemma it follows that, for any γ in $\bar{\Omega} \times \bar{\Gamma}$,

$$Ev \leqq Ee^v \leqq 1. \tag{2.7}$$

If v is $-\infty$ with probability one according to $f(x \mid \gamma_0)$, then (2.6) is obvious. Suppose therefore that $v > -\infty$ with positive probability according to

$f(x \mid \gamma_0)$. Then, by Jensen's inequality and (2.7),

$$Ev \leqq \log Ee^v \leqq 0, \tag{2.8}$$

and the first equality sign can hold only if v is a constant c with probability one according to $f(x \mid \gamma_0)$. If the first equality sign does not hold (2.6) follows at once. Consider, therefore, the constant c. If $c < 0$ then (2.6) holds. If $c > 0$ then (2.8) is violated. We cannot have $c = 0$ because of Assumption 4. This proves (2.6).

Now, as $\rho \downarrow 0$, for $\gamma' \neq \gamma_0$,

$$\lim E\left[\log \frac{w(X \mid \gamma, \rho)}{f(X \mid \gamma_0)}\right]^+ = E\left[\log \frac{f(X \mid \gamma)}{f(X \mid \gamma_0)}\right]^+ \tag{2.9}$$

by (2.3), (2.5), and Lebesgue's dominated convergence theorem. Also,

$$\lim E\left[\log \frac{w(X \mid \gamma, \rho)}{f(X \mid \gamma_0)}\right]^- = E\left[\log \frac{f(X \mid \gamma)}{f(X \mid \gamma_0)}\right]^-, \tag{2.10}$$

since the integrand of the left member decreases monotonically to the integrand of the right member. Hence, as $\rho \to 0$,

$$\lim E\left[\log \frac{w(X \mid \gamma, \rho)}{f(X \mid \gamma_0)}\right] = E \log \frac{f(X \mid \gamma)}{f(X \mid \gamma_0)} < 0 \tag{2.11}$$

by (2.6). Just as in [5] (see also [12]) it may then be shown that, for any positive ρ, there exists an $h(\rho), 0 < h(\rho) < 1$, such that the probability is one that, for all n sufficiently large,

$$\sup \left\{ \frac{\prod_{i=1}^n f(X_i \mid \gamma)}{\prod_{i=1}^n f(X_i \mid \gamma_0)} \right\} < h^n, \tag{2.12}$$

the supremum being taken over all γ in $\bar{\Omega} \times \bar{\Gamma}$ for which $\delta(\gamma, \gamma_0) > \rho$, and where $X_1, X_2, \cdots$ are independent chance variables with the common density $f(x \mid \gamma_0)$.

Let $L(x_1, \cdots, x_n \mid \gamma) = \prod_1^n f(x_i \mid \gamma)$. A *modified m.l. estimator* is defined to be a sequence of μ-measurable functions $\{\hat{\gamma}_n\}$ such that

$$L(x_1, \cdots, x_n \mid \hat{\gamma}_n(x_1, \cdots, x_n)) \geqq c \sup_\gamma L(x_1, \cdots, x_n \mid \gamma)$$

for almost all (μ) $x_1, \cdots, x_n$ for each n, where c is a positive number (the supremem is over $\Omega \times \Gamma$); for $c = 1$, this of course defines an m.l. estimator. (We shall not be concerned in this paper with conditions which ensure the existence of such measurable functions, although reasonable conditions are not difficult to formulate.) We also define a *neighborhood m.l. estimator* to be a sequence of μ-measurable functions $\{\gamma_n^*\}$ such that there exists a sequence of positive numbers ϵ_n with $\lim_{n\to\infty} \epsilon_n = 0$ for which $\sup_{\gamma \in \Pi_n} L(x_1, \cdots, x_n \mid \gamma) = \sup_\gamma L(x_1, \cdots, x_n \mid \gamma)$ for almost all $(\mu) x_1, \cdots, x_n$, where Π_n is the set of all γ in $\Omega \times \Gamma$ for which

$\delta(\gamma, \gamma_n^*(x_1, \cdots, x_n)) < \epsilon_n$. (Thus, neighborhood m.l. estimators exist in some cases where m.l. and modified m.l. estimators do not; this will be useful in making clear certain examples below where the lack of consistency is not merely due, as it might seem, to the fact that no strict m.l. or modified m.l. estimator exists.)

The above result (2.12) implies the strong convergence of m.l., modified m.l., and neighborhood m.l. estimators (in the respective cases where they exist). The component of the estimator which estimates G_0 converges to it at all its points of continuity w.p.1.

We remark that the above proof actually demonstrates consistency if, in the definition of m.l. estimator (or its variants), the supremum is taken over $\bar{\Omega} \times \bar{\Gamma}$ instead of over $\Omega \times \Gamma$ or, in fact, over any subset of $\bar{\Omega} \times \bar{\Gamma}$ containing γ_0. This last fact implies that if consistency is verified in an example where $\Omega = \Omega_1$, $\Gamma = \Gamma_1$, then it automatically holds in the example where $\Omega = \Omega_2$, $\Gamma = \Gamma_2$, whenever $\Omega_2 \subset \Omega_1$ and $\Gamma_2 \subset \Gamma_1$. In particular, this remark applies to the examples of Sections 3, 4, and 5.

We remark that Assumption 1 is not really essential in the above proof. Let P_γ denote the probability measure of X when γ is the true parameter value, and let $d(x, \gamma, \gamma_0) = r(x, \gamma, \gamma_0)/[1 - r(x, \gamma, \gamma_0)]$, where $r(x, \gamma, \gamma_0)$ denotes a Radon-Nikodym derivative of P_γ with respect to $P_\gamma + P_{\gamma_0}$ at the point x. If, for each $\gamma_0 \varepsilon \Omega \times \Gamma$, Assumptions 2 and 3 are satisfied when $f(x \mid \gamma)$ is replaced by $d(x, \gamma, \gamma_0)$, if (2.4) is replaced by the condition that $d(x, \gamma, \gamma_0) = 1$ does not hold on a set of probability one under γ_0 for any γ, and if $f(x \mid \gamma)/f(x \mid \gamma_0)$ is replaced by $d(x, \gamma, \gamma_0)$ (with a similar replacement for $w(x \mid \gamma, \gamma_0)$) in Assumption 5 and in the argument of the section, then (2.12) (with the replacement noted above) will still hold. An m.l. estimator $\hat{\gamma}$ is now defined to be one for which $\sup_\gamma \prod_1^n d(X_i, \gamma, \hat{\gamma}) = 1$ (with an analogous definition of modified and neighborhood m.l. estimator). We have not stated our assumptions and result (2.12) in this more general setting above because the stated form of the assumptions will suffice in most applications and will be easier to verify than assumptions stated in terms of $d(x, \gamma, \gamma_0)$ (which must be verified for each γ_0). As an example of the use of the more general result just cited, consider the problem of estimating the df F of a sequence of independent identically distributed discrete random variables, it being assumed that the true probability measure P_F (corresponding to the df F) satisfies

$$\sum_x P_F(x) \log P_F(x) > -\infty,$$

the sum being over all points x for which $P_F(x) > 0$. Then the assumptions are easily seen to be satisfied, and we may conclude that the sample df, which is the m.l. estimator, is a consistent estimator of F, a well-known result which does not usually seem to be considered as an example of the consistency of the m.l. estimator. (Of course, even if no restrictions of discreteness or logarithmic summability are placed on P_F, the sample df is still consistent and, as pointed

out in the introduction, this is the m.l. estimator. However, Assumption 5 is not satisfied in this case.)

Before proceeding to our examples in subsequent sections, we prove a simple lemma which will be useful later in verifying Assumption 5.

LEMMA. *If $f(z_1, \cdots, z_k)$ is a bounded probability density function with respect to Lebesgue measure μ on Euclidean k-space R^k, and if*

$$\int_{|z_i|>1} (\log |z_i|) f \, d\mu < \infty \qquad (1 \leqq i \leqq k), \tag{2.13}$$

then

$$-\int_{R^k} f \log f \, d\mu < \infty. \tag{2.14}$$

PROOF: If we prove that (2.13) implies (2.14) when f is replaced by cf in these equations, where $c > 0$, then the lemma is clearly proved. Thus, since f was assumed bounded, we may hereafter assume $f \leqq (2e)^{-1}$. (The new f need not have integral unity.) Let

$$g(z_1, \cdots, z_k) = f(z_1, \cdots, z_k) + \prod_{i=1}^{k} (z_i^2 + 1)^{-1}. \tag{2.15}$$

Clearly, (2.13) is true with f replaced by g. Moreover, since $g(z_1, \cdots, z_k) < e^{-1}$ outside of a sufficiently large sphere about the origin, and since $-f \log f < -g \cdot \log g$ if $0 < f < g < e^{-1}$, it suffices to prove (2.14) with f replaced by g, assuming g bounded and (2.13) with f replaced by g. By (2.13), we have

$$\int_{R^k} g \log \prod_{i=1}^{k} (1 + z_i^2)^{\frac{1}{2}} \, d\mu < \infty. \tag{2.16}$$

Thus, it suffices to prove the finiteness of

$$\begin{aligned} &-\int_{R^k} g \log g \, d\mu - \int_{R^k} g \log \prod_{i=1}^{k} (1 + z_i^2)^{\frac{1}{2}} \, d\mu \\ &\quad = \int_{R^k} g \log \prod_{i=1}^{k} (1 + z_i^2)^{\frac{1}{2}} \left\{ \frac{-\log [g \prod (1 + z_i^2)^{\frac{1}{2}}]}{\log \prod (1 + z_i^2)^{\frac{1}{2}}} \right\} d\mu. \end{aligned} \tag{2.17}$$

The fact that $g(z_1, \cdots, z_k) \geqq \prod (z_i^2 + 1)^{-1}$ (see (2.15)) implies easily that the bracketed expression in (2.17) is $\leqq 1$; by (2.16), this completes the proof of the lemma.

3. Example 1. Structural location parameter, incidental scale parameter. Let k be a positive integer, let μ be Lebesgue measure on Euclidean k-space, let g be a univariate probability density function with respect to Lebesgue measure, and let

$$f(x_i \mid \theta, \alpha_i) = \frac{1}{\alpha_i^k} \prod_{j=1}^{k} g\left(\frac{x_{ij} - \theta}{\alpha_i}\right), \tag{3.1}$$

where $x_i = (x_{i1}, \cdots, x_{ik})$. (Thus, observations are taken in groups of $k \geqq 1$, the value of the incidental parameter being the same within each group. The (unconditional) density of $X_i = (X_{i1}, \cdots, X_{ik})$ is given by Equation (2.1). Thus, the X_i are independent, but, for fixed i, the $X_{ij} (j = 1, \cdots, k)$ need not be independent.) Here Ω is the real line. Some further assumptions on g will be made below; we remark here that the important case

$$g(x) = (2\pi)^{-\frac{1}{2}} e^{-(x^2/2)} \tag{3.2}$$

will satisfy our assumptions. (See also (3.4) below.)

The cases $k = 1$ and $k > 1$ are essentially different. In Example 1a the consistency of the m.l. estimator will be proved for $k = 1$ assuming that A is the set of values $\alpha \geqq c$ where c is a known positive constant, and it is pointed out that the property of consistency of the m.l. estimator *does not hold* without this assumption. The proof of consistency in Example 1a is actually carried out for $k \geqq 1$ since this requires little additional space and will save space in Example 1b where we may refer back to 1a for proofs. In Example 1b we prove consistency of the m.l. estimator in the case $k > 1$ without assuming $\alpha \geqq c > 0$.

Example 1a. We assume that $k \geqq 1$ and that A is the set of all real values $\alpha \geqq c$ where c is a known *positive* constant. In the case $k = 1$, this assumption on A can be weakened slightly to an assumption on the behavior of $G(\alpha)$ as $\alpha \to 0$; however, some such assumption is necessary for consistency, since the last example of Section 6 shows that, even in cases where Γ is restricted to a simple parametric class of df's on a set of positive reals which is *not* bounded away from zero, it can happen that no m.l. or modified m.l. estimator exists and that there are neighborhood m.l. estimators which are not consistent.

We now state our assumptions on g and G_0. They seem reasonable and are in a form which makes brief proofs possible; they undoubtedly can be weakened. (These last remarks apply also to Examples 2 and 3. See also the first part of Section 6 for one method by which we can prove the results of our examples under weaker conditions.) We hereafter assume

(a) $\sup_x g(x) < \infty$;

(b) g is lower semicontinuous and for every $\epsilon > 0$ there is a continuous function $h_\epsilon \geqq g$ for which $\int [h_\epsilon(x) - g(x)]\, dx < \epsilon$;

(c) $\lim_{|x| \to \infty} g(x) = 0$;

(d) $$-\int_{-\infty}^{\infty} g(x)[\log |x|]^+ dx > -\infty; \tag{3.3}$$

(e) $$\int_{-\infty}^{\infty} |x|^{it} g(x)\, dx \neq 0 \quad \text{for almost all real } t;$$

(f) $g(x) > 0$ for almost all x in some open interval whose closure contains the point $x = 0$.

We note that, in addition to being satisfied in the case (3.2), Assumption (3.3) is also satisfied in such important cases as

(3.4) (a) $g(x) = 1/\pi(1 + x^2)$;
(b) $g(x) = 1$ if $|x| < \frac{1}{2}$ and $g(x) = 0$ otherwise;
(c) $g(x) = e^{-x}$ if $x > 0$ and $g(x) = 0$ otherwise.

Of course, if g does not satisfy (3.3) but if there is a function g^* satisfying (3.3) and for which $g(x) = g^*(x)$ almost everywhere, then we may carry out our considerations replacing g by g^*.

We assume that Γ consists of all G such that

$$\int_c^\infty (\log \alpha)\, dG(\alpha) < \infty, \tag{3.5}$$

where c is the constant used before in the definition of A. For example, G belongs to Γ if, for some positive constants b and ϵ,

$$1 - G(\alpha) < \frac{b}{\log \alpha (\log\log \alpha)^{1+\epsilon}} \tag{3.6}$$

for $\alpha > e^e$; integration by parts will verify that (3.6) implies (3.5). Condition (3.5) is weaker than the requirement that any positive (not necessarily integral) movement of G be finite.

We now verify the assumptions of Section 2. We complete the definition of f for (θ, α) in $\bar{\Omega} \times \bar{A}$ by setting $f(x \mid \theta, \alpha) = 0$ whenever $\theta = \pm\infty$ or $\alpha = \infty$. For $(\theta, G) \varepsilon\, \bar{\Omega} \times \bar{\Gamma}$, we then define $f(x \mid \theta, G)$ by (2.1). (We remark that $\bar{\Gamma}$ obviously contains all df's on $\bar{A}$.) Assumption 1 is obviously satisfied. Assumption 3 follows from the fact that (3.3) implies that $f(x \mid \theta, G)$ is for each x lower semicontinuous in (θ, G) (in the sense of the metric δ) on $\bar{\Omega} \times \bar{\Gamma}$, and the fact that $\bar{\Omega} \times \bar{\Gamma}$ is separable. Write $h_\epsilon(x_i \mid \theta, \alpha) = \alpha^{-k} \prod_{j=1}^k h_\epsilon[(x_{ij} - \theta)/\alpha]$. In order to verify Assumption 2, we note that, by the lower semicontinuity in (θ, G) of $f(x \mid \theta, G)$ and by the Helly-Bray theorem, we have (assuming, as we may, that the h_ϵ of (3.3) (b) satisfies $\lim_{|x|\to\infty} h_\epsilon(x) = 0$) that $(\theta_i, G_i) \to (\theta^*, G^*)$ as $i \to \infty$ implies

$$\begin{aligned} f(x \mid \theta^*, G^*) &\leqq \liminf_{i\to\infty} \int f(x \mid \theta_i, \alpha)\, dG_i \leqq \limsup_{i\to\infty} \int f(x \mid \theta_i, \alpha)\, dG_i \\ &\leqq \lim_{i\to\infty} \int h_\epsilon(x \mid \theta_i, \alpha)\, dG_i = \int h_\epsilon(x \mid \theta^*, \alpha)\, dG^*. \end{aligned} \tag{3.7}$$

Since the last member of (3.7) is greater than or equal to the first for all x and since their difference has integral $< \epsilon^k$ (with respect to μ), Assumption 2 follows at once.

In verifying Assumption 4, it clearly suffices to prove that, if $f(x \mid \theta_0, G_0) =$

$f(x \mid \theta_1, G_1)$ for almost all x, where $(\theta_i, G_i) \varepsilon \Omega \times \Gamma$ for $i = 0, 1$, then $(\theta_0, G_0) = (\theta_1, G_1)$. If an interval $0 < x < \epsilon$ satisfies (3.3) (f), there is a value β such that

$$P\{X_{1j} \leqq t \text{ for } 1 \leqq j \leqq k \mid \theta_0, G_0\} \leqq \beta$$

is satisfied (whatever be G_0) if and only if $t \leqq \theta_0$, a similar assertion holding if the interval $-\epsilon < x < 0$ satisfies (3.3) (f). Hence, it suffices to prove the above assertion when $\theta_0 = \theta_1$, since it cannot hold when $\theta_0 \neq \theta_1$. Let H_i be the df of the random variable $\log \alpha_1$ when G_i is the df of the random variable α_1; i.e., $H_i(t) = G_i(e^t)$. Then, putting $g^*(z) = e^z[g(e^z) + g(-e^z)]$ (g^* is the density of $\log |U|$ when g is the density of U), it suffices to prove that, if H_0 and H_1 are not identical, then $p_1(z_1, \cdots, z_k)$ and $p_2(z_1, \cdots, z_k)$ are not identical for almost all $(z_1, \cdots, z_k)$, where

$$p_i(z_1, \cdots, z_k) = \int_{-\infty}^{\infty} \prod_{j=1}^{k} g^*(z_j - \beta)\, dH_i(\beta). \tag{3.8}$$

Let g^{**} be the density function of $\sum_1^k Z_j/k$ when the Z_j are independent random variables with common density g^*. The above assertion is then implied by the assertion that the function

$$q(r) = \int_{-\infty}^{\infty} g^{**}(r - \beta)\, dH(\beta) \tag{3.9}$$

uniquely determines the df H. But if A, B, C are the characteristic functions of q, g^{**}, H, respectively, then $B(t) \neq 0$ for almost all t by (3.3) (e) and hence $C(t)$ is determined for those t for which $B(t) \neq 0$ by $C(t) = A(t)/B(t)$ and elsewhere by continuity. Thus, Assumption 4 is verified.

It remains to verify Assumption 5. Since $f(x \mid \theta, G)$ is uniformly bounded in x, θ, G, Assumption 5 will clearly be satisfied if

$$E \log f(X_1 \mid \theta_0, G_0) > -\infty. \tag{3.10}$$

Since the left side of (3.10) does not depend on θ_0, we may assume $\theta_0 = 0$. By (3.3) (d) and (3.5), we have

$$\begin{aligned} E[\log \mid X_{11} \mid]^+ &= E\left[\log \frac{\mid X_{11} \mid}{\alpha_1} + \log \alpha_1\right]^+ \\ &\leqq E\left[\log \frac{\mid X_{11} \mid}{\alpha_1}\right]^+ + E[\log \alpha_1]^+ < \infty; \end{aligned} \tag{3.11}$$

equation (3.10) is a consequence of (3.11) and the lemma at the end of Section 2.

This completes our verification of the fact that the assumptions of Section 2 are implied by (3.3) and (3.5).

Example 1b. We now assume $k > 1$. A is the set of all positive α, while Γ is the set of all df's G on A satisfying

$$\int_0^{\infty} \mid \log \alpha \mid dG(\alpha) < \infty. \tag{3.12}$$

We assume that g satisfies (3.3) (some alterations could be made here but, for the sake of brevity, we forego making them) and also that

$$\text{(3.13)} \qquad \text{(a)} \quad \lim_{|x|\to\infty} xg(x) = 0;$$

$$\text{(b)} \quad \sup_{x_1} [\min_{r<j} | x_{1r} - x_{1j} |]^k \prod_{j=1}^{k} g(x_{1j}) < \infty.$$

Assumption (3.13) is easily verified, for example, in cases (3.2) and (3.4).

We now verify the assumptions of Section 2. We define $f(x \mid \theta, \alpha) = 0$ whenever $\theta = \pm\infty$ or $\alpha = 0$ or ∞; $f(x \mid \theta, G)$ is then defined by (2.1) for $(\theta, G) \varepsilon \bar{\Omega} \times \bar{\Gamma}$. Assumptions 1, 3, and 4 are verified exactly as in Example 1a. In verifying Assumption 2, we may follow the demonstration of Example 1a, noting only that the h_ϵ of (3.3) (b) may (because of (3.13) (a)) clearly be assumed to satisfy $\lim_{|x|\to\infty} xh_\epsilon(x) = 0$, so that for every x none of whose components is θ^*,

$$\text{(3.14)} \qquad \lim_{\substack{i\to\infty \\ \alpha\to 0}} h_\epsilon(x \mid \theta_i, \alpha) = 0;$$

thus, for almost all (μ) x, the Helly-Bray theorem may still be used at the last step of (3.7), no difficulty being caused by the possibility that $\liminf_{i\to\infty} G_i(0) < G^*(0)$.

It remains to verify Assumption 5. Now, $f(x \mid \theta, G)$ is no longer uniformly bounded as it was in Example 1a. However, by (3.13) (b), there is a constant B such that, for all $x_1 = (x_{11}, \cdots, x_{1k})$ none of whose components are equal, every $\theta \varepsilon \Omega$, and every $\alpha \varepsilon A$,

$$\text{(3.15)} \qquad \begin{aligned} f(x_1 \mid \theta, \alpha) &= [\min_{r<s} | x_{1r} - x_{1s} |]^{-k} \left\{ [\min_{r<s} | y_{1r} - y_{1s} |]^k \prod_{j=1}^{k} g(y_j) \right\} \\ &\leqq B[\min_{r<s} | x_{1r} - x_{1s} |]^{-k}, \end{aligned}$$

where $y_{1r} = (x_{1r} - \theta)/\alpha$. Hence, for almost all x_1,

$$\text{(3.16)} \qquad \begin{aligned} \sup_{\substack{\theta\varepsilon\Omega \\ \alpha\varepsilon A}} \log f(x_1 \mid \theta, \alpha) &\leqq \log B + k \max_{r<s} \log [1/ | x_{1r} - x_{1s} |] \\ &\leqq \log B + k \sum_{\substack{r,s \\ r<s}} [\log (1/| x_{1r} - x_{1s} |)]^+. \end{aligned}$$

Now, by (3.3) (a), there is a value B' such that $g(z) \leqq B'$ for all z. Hence, by (3.12), B_1 denoting a finite constant, we have

$$\text{(3.17)} \qquad \begin{aligned} E[\log(1/ | X_{11} - X_{12} |)]^+ &\leqq E[\log 1/\alpha_1]^+ + E[\log (\alpha_1/ | X_{11} - X_{12} |)]^+ \\ &\leqq B_1 - 2 \int_{-\infty}^{\infty} g(z_2) \int_{z_2}^{z_2+1} B' \log (z_1 - z_2)\, dz_1\, dz_2 \\ &= B_1 + 2B' < \infty. \end{aligned}$$

From (3.16) and (3.17), we obtain

$$E \sup_{\gamma \in \bar{\Omega} \times \bar{\Gamma}} \log f(X_1 \mid \gamma) < \infty. \tag{3.18}$$

Assumption 5 is a consequence of (3.18) and of (3.10), the latter of which is proved exactly as in Example 1a. This completes the verification of the assumptions of Section 2 in Example 1b.

The discrete analogue of Example 1 can be carried out similarly by letting x, θ, α take on only rational values; this is, however, of less practical importance. The multivariate extension of Example 1 (X_{ij} a vector) may also be carried out similarly.

4. Example 2. The straight line with both variables subject to error.

In this section we shall treat the case $k = 1$ of fitting a straight line with both variables subject to normal error, a famous problem with a long history.

We consider a system $\{(X_{i1}, X_{i2})\}$, $i = 1, 2, \cdots$, of independent chance 2-vectors (the two components X_{i1}, X_{i2} need not be independent for fixed i). We have $\theta = (\theta_1, \theta_2)$, Ω the entire plane, $\theta_0 = (\theta_{10}, \theta_{20})$, A the entire line. Γ is the totality of all non-normal (univariate) distributions G (a chance variable which is constant with probability one is to be considered normally distributed with variance zero) which satisfy

$$\int (\log \mid \alpha \mid)^+ dG(\alpha) < \infty.$$

It is known to the statistician that

$$X_{i1} = \alpha_i + u_i,$$

$$X_{i2} = \theta_{10} + \theta_{20}\alpha_i + v_i,$$

where (u_i, v_i) are jointly normally distributed chance variables with means zero, each pair (u_i, v_i) distributed independently of every other pair and of the independent chance variables $\{\alpha_i\}$, with a common covariance matrix which is unknown to the statistician.

It is known (see [10]) that the distribution of (X_{i1}, X_{i2}) then determines θ_0 uniquely, but in general not G_0, the "true" df of α_i, or the "true" covariance matrix

$$\begin{Bmatrix} d_{11}^0 & d_{12}^0 \\ d_{12}^0 & d_{22}^0 \end{Bmatrix}$$

of (u_i, v_i). However, a "canonical" complex is determined. (See [4].)

Complete the spaces Ω, A, and Γ to obtain $\bar{\Omega}$, $\bar{A}$ and $\bar{\Gamma}$. The space $\bar{\Gamma}$ contains all normal distributions on A, but this will cause us no trouble in estimating θ_0, as we shall soon see.

Let D be the space of all triples (d_{11}, d_{12}, d_{22}) such that

$$d_{11} \geqq \lambda_{11} > 0, \qquad d_{22} \geqq \lambda_{22} > 0,$$
$$d_{11} d_{22} - d_{12}^2 \geqq \lambda_{12} > 0,$$

where $\lambda_{11}, \lambda_{12}, \lambda_{22}$, are given positive numbers. (This will be discussed further below.) We define a metric in D in the same way that one is defined on Ω. Let $\bar{D}$ be the completed space. We shall assume that the "true" triple $d_{11}^0, d_{12}^0, d_{22}^0$ is in D.

The place of $\bar{\Omega} \times \bar{\Gamma}$ in Section 2 and in Example 1 will now be taken by $\bar{\Omega} \times \bar{\Gamma} \times \bar{D}$. We therefore define

$$\gamma = (\theta_1, \theta_2, G, d_{11}, d_{12}, d_{22})$$

as the generic point in $\bar{\Omega} \times \bar{\Gamma} \times \bar{D}$.

Let $f(x_1, x_2 \mid \theta_1, \theta_2, \alpha, d_{11}, d_{12}, d_{22})$ be the joint density function of (X_{i1}, X_{i2}) when $\theta = (\theta_1, \theta_2)$, $\alpha_i = \alpha$, and the covariance matrix of (u_i, v_i) is

$$\begin{Bmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{Bmatrix}$$

(μ is Lebesgue measure in the plane). If, in the above, θ is in $\bar{\Omega} - \Omega$ or α is in $\bar{A} - A$ or (d_{11}, d_{12}, d_{22}) is in $\bar{D} - D$, we define f to be zero. Finally we define

$$f(x_1, x_2 \mid \gamma) = \int_{\bar{A}} f(x_1, x_2 \mid \theta_1, \theta_2, \alpha, d_{11}, d_{12}, d_{12})\, dG(\alpha).$$

It is known ([10] and [4]) that all γ in the same canonical class, and only such, define the same $f(x_1, x_2 \mid \gamma)$ (of course, to within a set of μ-measure zero). Two members of the same canonical class have the same $\theta = (\theta_1, \theta_2)$ but different G's and d_{ij}'s. We shall estimate only θ_0. For an estimator of the entire canonical complex by the minimum distance method under necessary assumptions only, see [4].[5] In Section 5 below will be found an explanation of why the entire canonical complex cannot be estimated by the m.l. method.

From the definition of $f(x_1, x_2 \mid \gamma)$ it follows immediately that Assumptions 1, 2, and 3 of Section 2 are satisfied. Since we are estimating only θ_0, it is sufficient to verify Assumption 4 only for θ_0 and $\theta^* \neq \theta_0$, i.e., if we write the γ_0 and γ_1 of (2.4) as

$$\gamma_0 = (\theta_{10}, \theta_{20}, G_0, d_{11}^0, d_{12}^0, d_{22}^0),$$
$$\gamma_1 = (\theta_1^*, \theta_2^*, G_1, d_{11}, d_{12}, d_{22}),$$

only $\theta_0 = (\theta_{10}, \theta_{20})$ has to be different from the corresponding $\theta^* = (\theta_1^*, \theta_2^*)$. Now we know that G_0 is in Γ, hence is not normal and assigns probability one to A. If G_1 is also in Γ then Assumption 4 follows at once from the results of

[5] See footnote 4.

Reiersøl [10] or from [11]. If G_1 assigns probability less than one to A, $f(x \mid \gamma_1)$ assigns probability less than one to the Euclidean plane of (x_1, x_2). If G_1 is normal and assigns probability one to A, then (X_{i1}, X_{i2}) are jointly normal under γ_1, but not under γ_0. Thus Assumption 4 is always satisfied.

To verify Assumption 5 we proceed essentially as in Example 1, and use the lemma at the end of Section 2. Assumption 5 is satisfied if

$$E \log f(X_{i1}, X_{i2} \mid \gamma_0) > -\infty.$$

By the lemma this will follow if we prove

$$E\{\log |X_{ij}|\}^+ < \infty$$

for $j = 1, 2$. Now

$$\begin{aligned} E\{\log |X_{i1}|\}^+ &\leqq E\{\log [\, |X_{i1} - \alpha_i| + |\alpha_i| \,]\}^+ \\ &\leqq E\{\log [\, |X_{i1} - \alpha_i| + 1]\} + E\{\log |\alpha_i|\}^+ \\ &= E\{\log [\, |u_i| + 1]\} + E\{\log |\alpha_i|\}^+ \\ &< \infty. \end{aligned}$$

Similarly,

$$\begin{aligned} E\{\log |X_{i2}|\}^+ &\leqq E\{\log [\, |X_{i2} - \theta_{10} - \theta_{20}\alpha_i| + |\theta_{10} + \theta_{20}\alpha_i| \,]\}^+ \\ &\leqq E \log [\, |X_{i2} - \theta_{10} - \theta_{20}\alpha_i| + 1] + E\{\log |\theta_{10} + \theta_{20}\alpha_i|\}^+ \\ &\leqq E \log [\, |v_i| + 1] + \{\log |\theta_{10}|\}^+ + \mathrm{E} \log [1 + |\theta_{20}\alpha_i| \,] \\ &< \infty. \end{aligned}$$

Thus we have shown, under our assumptions on Γ and D, that Assumptions 1 through 5 of Section 2 are satisfied, so that the m.l. estimator of θ_0 converges strongly to θ_0 as $n \to \infty$.

The assumption on D (that d_{11}, d_{22}, and $d_{11}\,d_{22} - d_{12}^2$ are bounded away from zero) cannot be entirely dispensed with. For if D consists of all triples for which d_{11}, d_{22}, and $d_{11}\,d_{22} - d_{12}^2$ are positive, if S_n is the sample df of $x_{11}, \cdots, x_{n1}$, and if $\hat{\gamma}_\epsilon$ is the complex $(0, 0, S_n, \epsilon, 0, \sum_1^n x_{i2}^2)$, then it is easily verified that $\lim_{\epsilon \to 0} L((x_{11}, x_{12}), \cdots, (x_{n1}, x_{n2}) \mid \hat{\gamma}_\epsilon) = \infty$; thus, no m.l. or modified m.l. estimator exists, and there are neighborhood m.l. estimators which are not consistent (for θ).

The case $k > 1$ is much simpler to treat than the above case. It is easy to see that then the covariance matrix of (u_i, v_i) is uniquely determined, and from this it follows easily that the whole complex γ is uniquely determined. The problem can be treated in a manner similar to that of Examples 1b and 3b.

The problem of this section with the distribution of (u_i, v_i) other than normal may also be treated by the m.l. method, as in Examples 1 and 3. The last paragraph of Section 3 applies also to the present example.

5. Example 3. Structural scale parameter, incidental location parameter. We consider here the case of a structural scale parameter and an incidental location parameter; this reverses the roles of the two parameters of Example 1. Thus, we suppose μ to be Lebesgue measure on R^k and

$$f(x_i \mid \theta, \alpha) = \frac{1}{\theta^k} \prod_{j=1}^{k} g\left(\frac{x_{ij} - \alpha}{\theta}\right). \tag{5.1}$$

The cases $k = 1$ and $k > 1$ are essentially different, and we consider them separately.

Example 3a. *The case* $k = 1$. This example is another simple one where no m.l. estimator is consistent, and also shows, in a simpler setting, why in Example 2 the m.l. method was incapable of estimating the components of the canonical complex other than θ. Since Example 3a is intended to illustrate the *failure* of the m.l. method in certain situations, we shall for simplicity assume that g is given by (3.2); examples with other g (e.g., (3.4)) may be treated similarly. Ω may be taken to be any specified set of positive numbers containing more than one point; for the sake of brevity, we assume that Ω contains its greatest lower bound c (say) (and thus, that $c > 0$), but it is easy to carry through a similar demonstration (with modified or neighborhood m.l. estimators in place of m.l. estimators) when $c \in \Omega$. Γ is taken to be the class of all df's G on the real line for which $\int [\log |\alpha|]^+ dG(\alpha) < \infty$ and such that G has no normal component; i.e., no G in Γ can be represented as the convolution of two df's, one of which is normal with positive variance. (Γ may be further restricted, e.g., by the condition that for each G there is a bounded set outside of which G has no variation.)

All assumptions of Section 2 are easily verified except Assumption 4; there is no difficulty of identifiability in $\Omega \times \Gamma$, but there clearly *is* in $\bar{\Omega} \times \bar{\Gamma}$. Consider now the expression

$$\prod_{i=1}^{n} \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{1}{2}} c} e^{-(1/2c^2)(x_i - s)^2} \, dM(s). \tag{5.2}$$

It is clear that the maximum of (5.2) with respect to M can be achieved only by an M which assigns probability one to the interval $(\min (x_1, \cdots, x_n), \max (x_1, \cdots, x_n))$ and hence which has no normal component. This discussion of the expression (5.2) shows that, for every n, any m.l. estimator (the fact that the maximum is attained is easily verified) of (θ, G) subject to our assumption $\theta \geqq c$ *always* estimates θ to be c. Thus, no m.l. estimator of (θ, G) is consistent (unless $\theta = c$).

To summarize the result of this example, then, the m.l. method is incapable of estimating consistently the normal component of the df of the sequence $\{X_i\}$ of independent identically distributed random variables because, in every neighborhood of a point (θ, G) with $\theta > c$, there are points with $\theta = c$ (and for which the likelihood is larger).

Let N_σ denote the normal df with mean 0 and variance σ^2, and let $H_1 * H_2$ denote the convolution of the two df's H_1 and H_2.

It is interesting to note that, without any assumption on Γ (except the necessary identifiability assumption that G_0 has no normal component), the minimum distance method is capable of estimating (θ_0, G_0) consistently [4]. The difficulty noted above for the m.l. estimator is avoided by noting the *rate* at which the sample df S_n converges to the df $N_{\theta_0} * G_0$ of X_1 and estimating θ_0 *not* by the value t for which $N_t * H$ is closest to S_n for some normal-free H (this would encounter the same difficulty as the m.l. estimator, since, the smaller t is taken, the closer can $N_t * H$ be made to approximate S_n), but as the *largest* value for which there is an $N_t * H$ suitably close to S_n ("suitably" is connected with the rate mentioned above.)

One could modify the example as considered above so as not to require G_0 to have no normal component, and try then to escape the difficulty of nonidentifiability by asking for an estimator of the canonical representation of (θ, G), this representation consisting of two df's, the normal and nonnormal components of $N_\theta * G$. The previous demonstration then shows that no m.l. estimator of the canonical representation estimates it consistently, and thus illustrates, in a simpler setting than that of Example 2 with $k = 1$, why the m.l. estimator could not be used in Example 2 to estimate the components of the canonical complex other than θ.

We remark that it is easy in many cases such as that of the present example to prove a result such as the one that, (t_n, H_n) denoting an m.l. estimator of (θ_0, G_0) after n observations, the df $N_{t_n} * H_n$ converges w.p.1 to $N_{\theta_0} * G_0$ as $n \to \infty$. Such a property is much weaker than that of the consistency of the m.l. estimator, and does not lie much deeper than the Glivenko-Cantelli theorem.

Example 3b. *The case* $k > 1$. We assume f to be given by (5.1) with $k > 1$. The function g is assumed to satisfy the conditions (*a*), (b), (c), and (d) of (3.3); conditions (a) and (b) of (3.13), and

$$\int_{-\infty}^{\infty} e^{itx} g(x)\, dx \neq 0 \quad \text{for almost all real } t. \tag{5.3}$$

(As in Example 1a, weaker conditions could be assumed here if we assumed also $\theta \geqq c > 0$; the above conditions are analogous to those of Example 1b.) Thus, for example, (3.2) and (3.4) satisfy these assumptions. Ω is the set of all values $\theta > 0$, while A is the real line and Γ is the set of all df's G on A for which

$$\int_{-\infty}^{\infty} [\log \mid \alpha \mid]^+ \, dG(\alpha) < \infty. \tag{5.4}$$

We now verify the assumptions of Section 2. We define $f(x \mid \theta, \alpha) = 0$ when $\theta = 0$ or ∞ or $\alpha = \pm\infty$. The definition of $f(x \mid \theta, G)$ for $(\theta, G) \varepsilon\, \bar{\Omega} \times \bar{\Gamma}$ is then given by (2.1). Assumptions 1, 2, and 3 are now verified as in Example 1b, interchanging the roles of θ and α in the latter (including the definition of $h_\epsilon(x \mid \theta, \alpha)$) and noting that (3.14)) still holds for almost all (μ) x, with this interchange. In

order to verify Assumption 4, we note, for $(\theta, G) \in \Omega \times \Gamma$, that θ is determined by the density function of $X_{11} - X_{12}$ and that, for almost all real t, the characteristic function of G is then given by $B(t/k, \cdots, t/k)/[C(\theta t/k)]^k$ where $B(t_1, \cdots, t_k)$ is the characteristic function of $X_{11}, \cdots, X_{1k}$ and $C(t)$ is the characteristic function of g.

Finally, Assumption 5 is a consequence of equation (3.18), which is proved in the present case exactly as in Example 1b (using (3.15), (3.16), and (3.17), with α_1 replaced by θ in the latter), and of equation (3.10) (with f defined by (5.1)). Equation (3.10) in the present example is a consequence of the lemma at the end of Section 2 and of

$$\begin{aligned} E\{\log |X_{11}|\}^+ &\leqq E\{\log [\, |X_{11} - \alpha_1| + |\alpha_1| \,]\}^+ \\ &\leqq E \log [\, |X_{11} - \alpha_1| + 1] + E\{\log |\alpha_1|\}^+ < \infty. \end{aligned} \tag{5.5}$$

This completes the verification of the assumptions of Section 2 in Example 3b. The last paragraph of Section 3 applies also to the present example.

6. The Classical case. Miscellaneous remarks. It does not seem to have been noticed in the literature that a simple device exists for proving consistency of the m.l. estimator in certain cases where the regularity conditions of published proofs fail. This device may be used in the case studied in the present paper (to prove consistency in the examples under weaker conditions than those stated) as well as in the classical parametric case. We now illustrate this device in an example of the latter case.

When Γ consists only of distributions which give probability one to a single point, the problem of the present paper becomes the classical problem of estimating the parameter θ and the parameter σ (say) to which G_0 gives probability one. If θ may be any real value and σ any positive value, then the function $(1/\sigma)g((x - \theta)/\sigma)$ of Section 3 does not satisfy Wald's integrability condition or the corresponding condition of any other published proof; one verifies easily that (2.5) is not satisfied for any point in the (θ, σ) half-plane which lies on the line $\sigma = 0$. (The line $\sigma = 0$ has to be added to Ω in the process of forming $\bar{\Omega}$. As in earlier sections, we assume the true σ_0 to be >0.) Often, however, when the observations are considered as if they were taken in groups of two or more, the integrability condition will be satisfied. Such is the case, for example, with the density function

$$\frac{1}{\pi} \frac{\sigma}{\sigma^2 + (x_1 - \theta)^2} \cdot \frac{1}{\pi} \frac{\sigma}{\sigma^2 + (x_2 - \theta)^2}$$

and the normal density function

$$\frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\left\{-\frac{1}{2}\frac{(x_1 - \theta)^2}{\sigma^2}\right\} \cdot \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\left\{-\frac{1}{2}\frac{(x_2 - \theta)^2}{\sigma^2}\right\}.$$

(Of course the estimator from the normal distribution is known to be consistent, but this does not alter the validity of the example.) In such cases it

follows from Wald's proof [5] (using the compactification device used above) or from the result of Example 1b that the m.l. sequence of estimators considered only after an even number of observations in consistent, and from this it is an easy matter to show that the entire m.l. sequence of estimators is consistent.

We shall now discuss the integrability conditions of [5] and of the present paper. The integrability condition (2.5) involves the difference of two logarithms; the integrability condition as given by Wald in [5] requires the finiteness of the expected value of each logarithm. The form (2.5) is satisfied whenever the condition of [5] is, and has one other advantage which we shall now illustrate by an example. Let the observed chance variable X have density function $\theta e^{-\theta x}$ for $x > 0$ and zero elsewhere. The parameter θ is unknown and Ω is the positive half-line, so that $\bar{\Omega}$ contains the point $\theta = 0$. One verifies easily that the condition of [5], and hence (2.5), are satisfied. Suppose now that, instead of observing X, one observes $Y = e^{(e^X)}$, which therefore has the density function

$$\frac{\theta}{x}(\log x)^{-\theta-1}$$

for $x > e$, and zero elsewhere. One readily verifies that, when $\theta < 1$,

$$E \log \left\{ \frac{\theta}{Y} (\log Y)^{-\theta-1} \right\} = -\infty,$$

so that the condition of [5] is not satisfied when $0 < \theta_0 < 1$. Thus, whether the condition of [5] is satisfied depends in this instance on whether one observes X or Y; this is an unfortunate circumstance, since the estimation problems are in simple correspondence. On the other hand, condition (2.5) is invariant under one-to-one transformation of the observed chance variable because the numerator and denominator of the ratio in (2.5) are multiplied by the same Jacobian. (In particular, therefore, the chance variable Y satisfies (2.5).)

Without resorting to artificial or pathologic examples as is sometimes done in the literature, it is still easy to give instances where the m.l. method does not give consistent estimators in the classical parametric case. For example, consider the density function

$$\frac{1}{2(2\pi)^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(x-\theta)^2\} + \frac{1}{2(2\pi)^{\frac{1}{2}}\sigma} \exp\left\{-\frac{1}{2}\frac{(x-\theta)^2}{\sigma^2}\right\}$$

of the sequence of independent and identically distributed chance variables $X_1, X_2, \cdots$. Here θ and σ are the unknown parameters, θ may be any real number and σ any positive number. It is easy to see that the supremum of the likelihood function is almost always infinite, no m.l. or modified m.l. estimator exists, and there are neighborhood m.l. estimators (where θ_0 is estimated by X_1, say) which are obviously not consistent.

REFERENCES

[1] J. NEYMAN AND E. L. SCOTT, "Consistent estimates based on partially consistent observations" *Econometrica*, Vol. 16 (1948), pp. 1–32.

[2] A. WALD, "The fitting of straight lines if both variables are subject to error," *Ann. Math. Stat.*, Vol. 11 (1940), pp. 284–300.

[3] A. DVORETZKY, J. KIEFER AND J. WOLFOWITZ, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 642–669.

[4] J. WOLFOWITZ, "Estimation of the components of stochastic structures," *Proc. Nat. Acad. Sci., U.S.A.*, Vol. 40, No. 7 (1954), pp. 602–606.

[5] A. WALD, "Note on the consistency of the maximum likelihood estimate," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 595–601.

[6] A. WALD, "Estimation of a parameter when the number of unknown parameters increases indefinitely with the number of observations," *Ann. Math. Stat.*, Vol. 19 (1948), pp. 220–227.

[7] J. NEYMAN, "Existence of consistent estimates of the directional parameter in a linear structural relation between two variables," *Ann. Math. Stat.* Vol. 22 (1951), pp. 497–512.

[8] J. NEYMAN AND E. L. SCOTT, "On certain methods of estimating the linear structural relation between two variables," *Ann. Math. Stat.* Vol. 22 (1951), pp. 352–361.

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

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

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

[12] J. WOLFOWITZ, "On Wald's proof of the consistency of the maximum likelihood estimate," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 602–603.