The Annals of Statistics
1982, Vol. 10, No. 3, 717–740

OPTIMAL ESTIMATION OF A GENERAL REGRESSION FUNCTION

BY P. W. MILLAR[1]

University of California

Let Θ be a convex subset of R^d, and for each $\theta \in \Theta$ let $F(\theta; dt)$ be a probability on the line. For any vector $a_n = (a_{n1}, a_{n2}, \cdots, a_{nn})$ where $a_{ni} \in \Theta$, let $X_{n1}, \cdots, X_{nn}$ be independent observations, the distribution of X_{ni} being $F(a_{ni}; dt)$. The main results give a method of estimating the unknown regression function a_n based on a minimum distance recipe. Under regularity assumptions, the proposed estimators are shown, in an appropriate framework, to be asymptotically normal, locally asymptotically minimax, and robust. The abstract results are illustrated by application to the linear model and to exponential response models. In general, nothing at all is assumed about the form of the regression function; accordingly, this forces the limiting normal distributions of the proposed estimators to be located on infinite dimensional linear spaces.

1. Introduction.

Let Θ be an open subinterval of R^1, and let $\{F(\theta; dx), \theta \in \Theta\}$ be a fixed family of probability measures on the line. If n is a positive integer, let $a_n = (a_{n1}, \cdots, a_{nn})$ be a vector in R^n with $a_{ni} \in \Theta$, and let $X_n = (X_{n1}, \cdots, X_{nn})$ be a random vector of independent random variables, where the distribution of X_{ni} is $F(a_{ni}; dx)$. This setup will be called a regression model, and the vector a_n a regression function. The regression function a_n may be entirely unknown, or else may be known to lie in some lower dimensional subset of R^n. With this understanding, the main contribution of this paper is a method of estimating a_n optimally under very minimal assumptions as to its whereabouts. In an appropriate framework, the proposed estimators turn out to be asymptotically normal and robust.

Before describing these results, let us look at two illustrations of this abstract regression model which are of paramount importance for applications: shift models and exponential models.

Shift models. Fix F, a distribution on the line, and define $F(\theta; dx) = F(dx - \theta)$ for $\theta \in R^1$. With this choice of parametric family, the model of the preceding paragraph may then be described in the following equivalent way. Let $Z_1, \cdots, Z_n$ be iid F. Then each X_{ni} may be written $X_{ni} = a_{ni} + Z_i$. If a_{ni} is known to have the form (say) $a_{ni} = a(i/n) + b$ with a, b unknown, and if F is $N(0, 1)$, then this is the familiar straight-line regression problem, and the allowable regression functions a_n belong to a two dimensional subspace of R_n. This paper, of course, provides a method of estimating a, b that is optimal—in fact, robust—in a sense described below. On the other hand, if F is still $N(0, 1)$ and nothing at all is known about a_n, then we can still provide an estimator of a_n with strong optimality properties. More illustrations are provided in Section 2.

Exponential models.. Here $F(\theta; dx)$ in standard form has a density of the type $\exp\{\theta q(x) + b(\theta)\}$ relative to some sigma finite measure ν; see Johansen (1979) for an

Received October 1980; revised January 1982.

[1] Research supported by NSF grant MCS 75-10376.

AMS 1980 *subject classifications.* Primary 62G05, 62J02; secondary 62G20, 62F12, 62F35.

Key words and phrases. Non-parametric regression, robust regression, local asymptotic minimax, minimum distance estimation, empirical distribution, Kiefer process, exponential family, location model, $\sqrt{n}$ consistency, Hilbert space, abstract Wiener space, asymptotic normality, infinite dimensional normal families.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.*
www.jstor.org

introduction to the theory of such families. The parameter set Θ is the interval $\{\theta: -\infty < b(\theta) < \infty\}$. Regression models developed for such families include the logit model and their generalization, the exponential response models; see Haberman (1977) for this, as well as for reference to their application in educational testing. In his interesting article, Haberman uses a maximum likelihood method to provide, under certain conditions, asymptotically normal estimates of real functionals of a_n, but no optimality is proved. Here, using a different method, we estimate the vector a_n itself; the estimate turns out to be asymptotically normal and asymptotically minimax. Of course, one can then employ standard weak convergence results to study the asymptotic distribution of estimators of functionals of a_n.

The estimators of a_n employed in this paper, being variants of a minimum distance recipe, are fairly easy to describe intuitively. For each vector a_n, define $\bar{F}(a_n; s, t) = n^{-1}\sum_{\iota \le ns} F(a_{ni}; t)$, where $F(\theta; t)$ is the cdf of the measure $F(\theta; dx)$. Define a two-parameter stochastic process $\hat{F}_n(s, t)$, analogous to the empirical cdf, by $\hat{F}_n(s, t) = n^{-1}\sum_{i \le ns} I_{(-\infty, t]}(X_{ni})$. If $|\cdot|_m$ is a norm on bounded functions of two variables, estimate a_n to be the vector for which $\bar{F}(a_n; s, t)$ is closest to $\hat{F}_n$ in the sense of the distance obtained from $|\cdot|_m$. Each vector a_n is identified with a point in an infinite dimensional linear space L which does not depend on n. When an appropriate metric is placed on L, the estimators just described turn out to be $n^{1/2}$ consistent, and under regularity assumptions on the family $\{F(\theta; \cdot)\}$ they converge, when suitably normalized, to a normal distribution on L. This normal distribution, of course, will in general be infinite dimensional.

Finally, these estimators of a_n are shown to be robust. This has the usual intuitive interpretation: namely, the estimators do not deteriorate when, because of data contamination, the distributions of the X_{ni} do not follow $F(a_{ni}; dx)$ precisely for any choice of a_n. Least squares estimates of a_n, which one might naturally use in some of the regression problems mentioned above, are not robust in this sense because even one outlying observation drastically affects the estimate. As in Millar (1981a), we give the intuitive notion of robustness a technical definition that is purely decision theoretic. To describe this, let $F(n, a_n; dx)$ denote the product measure of $F(a_{ni}; dx_i)$. If ℓ_n is, for each n, a Hilbertian distance on measures (see Section 9 for the precise description), let $\ell_n(F(n, a_n), Q)$ be the loss if a_n is our estimate when the actual distribution of the data is given by the product measure Q. Let T_n denote an estimator of the regression function. A sequence of estimators $\{T_n^0\}$ will be called robust if

$$\lim_n \inf_{T_n} \sup_{Q \in C_n} \int \ell_n(F(n, T_n), Q)\, dQ = \lim_n \sup_{Q \in C_n} \int \ell_n(F(n, T_n^0), Q)\, dQ,$$

where $\{C_n\}$ is a sequence of neighborhoods of an arbitrary but fixed $F(n, a_n)$ shrinking at a certain rate as n increases; see Section 9 for the precise description. In short, an estimate is robust if it is locally asymptotically minimax in a certain decision theoretic framework. That such a framework convincingly captures the intent of robustness is discussed at length in Millar (1981a) for the location-scale problem (among others), and the arguments will not be repeated here; in the location problem (a special case of a regression shift model), the theories of this paper and the earlier one produce equivalent optimal procedures. The asymptotic minimax approach to robustness is discussed also by Beran (1979), Bickel (1979), and Rieder (1979), with different choices of metrics and neighborhoods.

The present paper differs from previous works on nonparametric and robust regression in several respects. First, previous efforts in robust regression have either restricted attention to (essentially) the case where a_n lies in a subset of R^n of fixed dimension d, or else, as in Huber (1973), one lets d increase with n but much more slowly. In the most general situation of this paper, there are no restrictions on a_n other than that its components belong to Θ; yet the proposed estimators will be robust and asymptotically normal when these notions are formulated properly for this situation. On the other hand, recent attacks on nonparametric regression (Stone, 1977; Sacks-Spiegelman, 1980) can provide consistent estimators in this generality; but to the best of my knowledge, this paper is the first to provide in this generality $\sqrt{n}$ consistency, asymptotic normality and asymptotic optimality.

The current theory of robust regression—such as it is—consists essentially in taking the basic least squares recipe and perturbing it slightly to make outliers less disruptive. This having been done, certain asymptotic normality results are proved (cf. Huber, 1973; Bickel, 1975; Maronna and Yohai, 1979). Unfortunately, ad hoc tinkering with a failing recipe to "save the phenomenon" is quite unsatisfying, and it seems clear that the theory of robust regression needs to be thoroughly overhauled. This paper offers one possibility, based on the concept of minimum distance and an abstract minimax principle. Grounds for believing that this might work are provided by the successes of these principles in the simpler problem of robust parametric estimation (see Beran, 1977, 1979; Millar, 1981a; Parr-Schucany, 1979; Rieder, 1979).

Sections 2, 3, and 4 describe the model more precisely, specify the objects to be estimated, and give the basic minimum distance estimators to be studied throughout. Sections 5 and 6 establish consistency and asymptotic normality, while Section 7 illustrates the application of the theoretical results to shift and exponential models. Robustness is established in Section 9, based on the theoretical preparation of Section 8. The final sections supply proofs of the main results.

The results of this paper are cast in a fairly abstract form. This was necessary because the basic problem here—being genuinely infinite dimensional—is inherently complicated. The applications sketched in Section 7 attempt to illustrate various concrete aspects of the theory. For even more explicit examples, see Millar (1981a), where explicit asymptotic forms of estimators like those of the present paper are developed for, e.g., the location model; indeed, a good understanding of the cited paper will make the reading of the present one much simpler. The rationale for the local asymptotic minimax approach to robustness is discussed in Millar (1981a) and Beran (1981). To understand the connection with classical parametric estimation, it is helpful to read Hajek (1972). A detailed development of the asymptotic minimax theory and many of its applications can be found in the lecture notes of Millar (1981b). Finally, it should be pointed out that the theory developed in the present paper is abstract enough so that further applications can be built upon it. These include (a) a theory of testing which will assess goodness of fit by means rather more effective than the usual examination of residuals, and (b) application to dimensionality reduction of data. These will be discussed elsewhere.

2. The model. This section gives a precise and mathematically convenient description of the regression models discussed in Section 1.

Fix a parametric family $\{F(\theta; dx), \theta \in \Theta\}$ where Θ is a subinterval of R^1. Denote by $F(\theta; t)$ the cdf of $F(\theta; dx)$. Let L denote the Hilbert space $L^2([0, 1], dx)$ where dx is Lebesgue measure, and set

$$|g|_L = \left(\int_0^1 g(s)^2\, ds\right)^{1/2}. \tag{2.1}$$

Define

$$c(n; i; g) = n \int_{(i-1)/n}^{i/n} g(s)\, ds. \tag{2.2}$$

Let Γ be a subset of L such that if $g \in \Gamma$ then range $g \subset \Theta$. One basic form of the regression model then consists in observing the vector $(X_{n1}, \cdots, X_{nn})$ of independent rv's, where X_{ni} has distribution $F(c(n; i; g); dx)$ for each i and some $g \in \Gamma$. The distribution of the vector, of course, is then given by the product measure

$$F(n; g; dx) = \textstyle\prod_i F(c(n; i; g); dx_i), \qquad x = (x_1, \cdots, x_n). \tag{2.3}$$

The set Γ specifies the allowable regression functions. Notice that if, say, Γ is dense in L then, despite the apparent restriction of the recipe (2.2), the allowable regression functions

$\{\{c(n; i; g)\}: g \in \Gamma\}$ are essentially all of R^n. The basic statistical problem is now to estimate the vector $\{c(n; i; g)\}$ or, equivalently, to pick the "right" $g \in \Gamma$.

Illustration: the shift model. As noted in Section 1, the observations for the shift model can be written $X_{ni} = c(n; i; g) + Z_i$ where Z_i are iid F, $g \in \Gamma$. Various choices of Γ lead to familiar statistical models. The simplest choice of Γ is the one dimensional subspace $\{\theta g_0\}$ where $g_0 \in L$ is fixed, and θ ranges over all real numbers. Estimation of $g \in \Gamma$ here of course is equivalent to estimating θ. If $g_0 \equiv 1$, this choice of Γ gives the familiar location model; if $g_0(s) = s$, $0 \le s \le 1$, the model is simple straight line regression through the origin. If $g_1, \cdots, g_k$ are fixed elements of L, then the choice $\Gamma = \{\sum_{i=1}^k \theta_i g_i : \theta_i \text{ real}\}$ leads to a model where the regression function is determined by only k parameters, and estimation of $g \in \Gamma$ here is tantamount to estimating $(\theta_1, \cdots, \theta_k)$.

If, in this case, g_i is the indicator of the interval from $(i - 1)/k$ to i/k and F_θ is the $N(\theta, 1)$ distribution, then the model is, essentially, the one-way layout with equal numbers of observations per category.

If Γ is the set of all elements of L that are step functions, this leads to a regression model in which the regression functions $\{\{c(n; i; g)\}g \in \Gamma\}$ are completely arbitrary points of R^n. As a somewhat different situation, take Γ to consist of all elements g such that $g(s) = 0$, $0 \le s < a$; $g(s) = b$, $a \le s \le 1$ for some $a \in (0, 1)$ and some real b. For this model, which is essentially a two sample problem, the first na observations follow F, but all the succeeding ones follow $F(t - b)$. Estimation of $g \in \Gamma$ is then tantamount to deciding when the shift in distribution occurred (a variant of the "earliest detection problem" familiar from sequential analysis) and how large the shift was when it occurred. Finally, it is perhaps helpful to note that, in the case of the shift model, the recipe (2.2) is a discrete variant of the stochastic differential equation $dX_t = g(t)\,dt + dB_t$ where $g \in L$ and B_t is standard Brownian motion. This equation is familiar from the engineering literature on signal detection, the statistical problem being that of estimating the signal g in the presence of "noise" dB_t.

It is convenient to give a different description of the regression coefficients $c(n; i; g)$. For this purpose, let

(2.4) $\mathscr{F}_n$ = sigma field on $[0, 1]$ generated by the subintervals $(i/n, i + 1/n]$, $0 \le i < n$.

If $g \in L$, define

$$T(n; g) = E(g \mid \mathscr{F}_n), \tag{2.5}$$

where this conditional expectation is computed using Lebesgue measure on $[0, 1]$. The value of $T(n; g)$ on the interval $(i - 1/n, i/n]$ is of course $c(n; i; g)$. Therefore the coefficients $c(n; i; g)$ are given by the successive values of the step function $T(n; g)(s)$ as s moves from 0 to 1. From now on we work directly with the functions $T(n; g)$; the effective parameter set at time n evidently is $\{T(n; g): g \in \Gamma\}$ and we take the basic statistical problem to be the optimal estimation of the functions $T(n; g)$.

Although the recipe (2.2) for the regression coefficients has a certain weight of tradition behind it (cf, Hájek, and Šidák, 1967), it is useful to consider variants. To this end, for integers $n \ge k$ define

$$T(n; k; g) = T(n; T(k; g)) = E[E\{g \mid \mathscr{F}_k\} \mid \mathscr{F}_n]. \tag{2.6}$$

If n is a multiple of k, then obviously $T(n; k; g) = T(k; g)$; otherwise, $T(n; k; g)$ will have at most $2k$ different values. Since $T(n; k; g)$ is still, by definition, a step function on the intervals $(i/n, i + 1/n]$ one can extend the regression model by specifying that the coefficients are, for specified integers $k_n \le n$, given by the n successive values of $T(n; k_n; g)$, $g \in \Gamma$. If $k_n = n$, the earlier model is recovered; if $k_n \ll n$, then the dimension of the parameter is far less than the number of observations. Accordingly, this last possibility can be viewed as a systematic way of "letting the dimension of the regression problem increase slowly with n." Alternatively, one could view it as one technical formulation allowing the grouping of data into blocks, the data being iid within each block.

We conclude this section with a simple convergence result that is used throughout the paper.

(2.7) PROPOSITION. *If $g \in L^p(0, 1), p \geq 1$, then $T(n; k_n; g)$ converges in L^p norm to g as $n \geq k_n \to \infty$.*

PROOF. Since conditional expectation is a contraction in L^p, the problem reduces easily to the case $k_n = n$. For this case, the proof is easily accomplished by a simple variant on the usual argument for the L^p convergence of martingales indexed by directed sets; see, for example, Chatterji (1973).

3. Metrics and reparametrization. The model delineated in Section 2 has, at time n, a parameter set which consists of functions $T(n; k_n; g)$, $g \in \Gamma$. The purpose of this section is to replace that parameter set by one that is much more convenient to deal with. To carry this out, define for $g \in \Gamma$ a function Fg on $[0, 1] \times R$ by

$$(Fg)(s, t) = \int_0^s F(g(u); t)\, du. \tag{3.1}$$

Sometimes g in (3.1) will have a complicated form. For example, $FT(n; k_n; g_0)$ is to mean Fg with $g = T(n; k_n; g_0)$. Assuming, as we do henceforth, that $F(\theta; t)$ is Borel measurable in θ, the recipe (3.1) makes sense and defines a distribution function on $[0, 1] \times R$. It is clear that Fg will be continuous if $F(\theta; t)$ is continuous for each θ. Assume further, throughout the rest of the paper, that

$$\text{if} \quad F(\theta; t) = F(\theta'; t) \quad \text{for all} \quad t, \qquad \text{then} \quad \theta = \theta'. \tag{3.2}$$

It is then immediate that if $Fg = Fh$ on a countable dense subset of $[0, 1] \times R$, then $g = h$ a.e. Accordingly, knowing the function Fg is tantamount to knowing g, and so we may take $\{FT(n; k_n; g) : g \in \Gamma\}$ as parameter set at time n instead of $\{T(n; k_n; g) : g \in \Gamma\}$. The basic statistical problem then becomes the estimation of the functions $FT(n; k_n; g)$, an estimate of which automatically produces an estimate of $T(n; k_n; g)$ and hence of the regression function. As we shall see, it is often not difficult to translate back from the new parameter set $\{FT(n; k_n; g) g \in \Gamma\}$ to the original one.

Next we shall introduce a metric for the new parameter set. To this end, let

$$m = \text{probability measure on } R^1 \text{ with support } \Theta. \tag{3.4}$$

Let L_m be the Hilbert space $L^2([0, 1] \times R, ds\, dm)$ and denote the norm of L_m by $|\cdot|_m$. If the metric obtained from $|\cdot|_m$ is put on the parameter set $\{FT(n; k_n; g)\}$, then of course it becomes a separable metric space.

While the preceding two paragraphs establish the parameter set with which we shall work most of the time, later applications require clarification of its ties to the original parameter set. Define a function d_F on $\Gamma \times \Gamma$ by

$$d_F(h, g) = |Fg - Fh|_m. \tag{3.5}$$

It is immediate that d_F is a metric on Γ. If $F(\theta; t)$ is continuous in θ, then (Γ, d_F) becomes a separable metric space; if g_n converges to g in the norm of L (or, more generally, if g_n converges to g in measure) then $d_F(g_n, g) \to 0$. Therefore, the metric d_F is weaker than the metric obtained from $|\cdot|_L$. The importance of this metric is (cf. Section 5) that there are $\sqrt{n}$ consistent estimators of the original parameters if the metric on L is d_F; this is false if the metric is given by the norm $|\cdot|_L$. The possibility of putting a norm $|\cdot|_0$ on L to replace d_F is considered in Section 6; this norm will be weaker than $|\cdot|_L$.

This section concludes with several propositions which, in special cases, characterize convergence in the sense of the metric d_F. These results are proved in Section 12. Recall that measurable functions g_n on the unit interval converge in measure to g if, for each ε, the Lebesgue measure of $\{t:|g_n(t) - g(t)| > \varepsilon\}$ goes to 0 as n increases.

(3.6) PROPOSITION. *Let $F(\theta; t) = F(t - \theta)$ where F is continuous and support $F = R^1$. For $g_n, g \in L$, $d_F(g_n, g) \to 0$ if and only if g_n converges to g in measure.*

(3.7) PROPOSITION. *Assume Γ is such that there exists a real number $b > 0$ and $|g|_L \le b, g \in \Gamma$. Assume $F(\theta; t)$ is continuous in θ and satisfies the condition*

(3.8) *for each fixed $s \in (0, 1)$ and $g, g' \in \Gamma$, if $Fg(s, t) = Fg'(s, t)$ for all t, then $g = g'$* a.e. *on $[0, s]$.*

Let $g_n, g \in \Gamma$. Then $d_F(g_n, g) \to 0$ if and only if g_n converges to g in measure.

(3.9) EXAMPLE. Consider the exponential family model introduced in Section 1. Using the uniqueness theorem for Laplace transforms, it is not difficult to check that hypothesis (3.8) holds, provided the support of the measure $A \to \nu\{x:q(x) \in A\}$ contains an interval.

The next proposition is trivial and will not be proved.

(3.10) PROPOSITION. *If Γ is a strongly compact subset of L, then for $g_n, g \in \Gamma$, $d_F(g_n, g) \to 0$ if and only if g_n converges to g in the norm of L.*

The typical situation in which (3.10) applies is when Γ is finite dimensional and norm bounded.

(3.11) PROPOSITION. *Suppose Γ is finite dimensional and for all θ, t, $0 < F(\theta; t) < 1$. Suppose $\Theta = (0, \infty)$ and $\lim_{\theta \to \infty} F(\theta; t) = 1$ for some t. Then $d_F(g_n, g) \to 0$ if and only if $g_n \to g$ in $|\cdot|_L$.*

This proposition applies to scale families: $F(\theta; t) = F(1; \theta t)$, $\theta > 0$. There are obvious perturbations of this result to parameter sets $(-\infty, \infty)$, $(-\infty, 0)$.

4. Minimum distance estimators. This section describes the estimators whose excellence is championed throughout the remainder of the paper. The parametric family $\{F(\theta; dx)\}$ and the measure m introduced in Sections 2 and 3 are henceforth fixed. To define the estimators, two new ingredients are needed. First, introduce a two parameter stochastic process $\hat{F}_n(s, t)$, which is to be the analogue for regression problems of the empirical cdf:

$$\hat{F}_n(s, t) = n^{-1} \textstyle\sum_{i \le ns} I_{(-\infty, t]}(x_{ni}), \qquad 0 \le s \le 1, \qquad -\infty < t < \infty. \tag{4.1}$$

Here, and throughout, I_A denotes the indicator function of the set A. For the model of Section 2, it is easy to see that, up to an error less than $O(n^{-1})$:

$$E\hat{F}_n = FT(n, k_n; g), \tag{4.2}$$

where expectation is computed relative to the product measure $F(n; T(n; k_n; g); dx)$ defined in (2.3). As to be expected, a suitable normalization of $\hat{F}_n$ converges to a two parameter Gaussian process; these matters are discussed in Section 5.

The second ingredient needed is a system $\{D_n\}$ of subsets of L. The systems D_n are subject to

$$D_n \subset D_{n+1}, \qquad n = 1, 2, \cdots. \tag{4.3}$$

Further hypotheses are imposed on D_n later, depending on the regression problem at hand. Examples of the kind of classes D_n to which the results will apply include: (a) $D_n = \Gamma$ a fixed subset of L, not necessarily proper; (b) $D_n = \operatorname{span}\{e_1, \cdots, e_{j_n}\}$, where $j_n \uparrow \infty$ and $\{e_i\}$ is an orthonormal set in L; (c) D_n = all step functions on the intervals

$\{(i/2^{a_n}, (i+1)/2^{a_n}]\}$, $a_n \uparrow \infty$; (d) D_n = all elements of L which have a first derivative bounded by M_n, where $M_n \uparrow \infty$. The subset $\Gamma \subset L$ giving the "allowable" g's for the regression problem will henceforth be assumed to satisfy $\Gamma = \cup D_n$. It is clear that if D_n is given by b, c, or d then $\Gamma = \cup D_n$ will be a subspace of L, dense in L but proper. For all practical purposes, one can regard Γ as L in these cases. The problem of deciding, say, whether $g \in L$ has a finite or infinite expansion in some given basis is not of compelling practical interest.

The minimum distance estimate $\hat{g}_n$ of $T(n; k_n; g)$, $g \in \cup D_n = \Gamma$, can now be defined to be any element of the form $T(n; k_n; h)$, $h \in D_n$, that satisfies

$$\inf_{g\in D_n} |\hat{F}_n - FT(n; k_n; g)|_m = |\hat{F}_n - F\hat{g}_n|_m$$

(or else comes within n^{-1} of achieving this inf—asymptotically it will not matter). This formulation does not use the reparametrization of Section 3. For that, define $\pi_n^0\hat{F}_n$ to be the element of the form $FT(n; k_n; h)$, $h \in D_n$, which achieves (or comes within n^{-1} of such)

$$\inf_{g\in D_n} |\hat{F}_n - FT(n; k_n; g)|_m = |\hat{F}_n - \pi_n^0\hat{F}_n|_m.$$

Assuming uniqueness, we have $F\hat{g}_n = \pi_n^0\hat{F}_n$. Of course, $\pi_n^0\hat{F}_n$ depends on the system $\{D_n\}$, but we suppress this in the notation.

In subsequent sections, these estimators will be shown to have desirable optimality properties. The basic results impose conditions on the systems $\{D_n\}$ as well as on the numbers k_n appearing in the regression functions $T(n; k_n; g)$. If one wants $\sqrt{n}$ consistency (cf. Section 5), then one can use estimators of this type with $D_n = \Gamma$ for all n. For the more delicate problem of asymptotic normality, it is necessary, if Γ is infinite dimensional, to consider systems $\{D_n\}$ that are more shrewdly chosen.

5. Weak convergence and consistency. This section introduces certain two-parameter Gaussian processes to which the sample cdf (cf. (4.1)) converges. As a by-product, the minimum distance estimators of Section 4 are shown to be $\sqrt{n}$ consistent, as defined in (5.4), in an appropriate framework. To lighten the notation, assume $k_n = n$ throughout.

Define a normalization of $\hat{F}_n$ by the recipe

$$\begin{aligned} W_n(s,t) = W_n(g; s,t) &= n^{1/2}\{\hat{F}_n(s,t) - E_{ng}\hat{F}_n(s,t)\} \\ &= n^{1/2}\{\hat{F}_n(s,t) - FT(n,g)(s,t)\} + O(n^{-1/2}), \end{aligned} \tag{5.1}$$

where E_{ng} is expectation under the product measure $F(n; T(n; g); dx)$ of (2.3). Though W_n depends on g, we shall usually suppress this in the notation. It is easy to see that W_n has mean 0 and covariance $K(T(n; g))(s_1, t_1; s_2, t_2)$ where

$$K(h)(s_1, t_1; s_2, t_2) = Fh(s_1 \wedge s_2; t_1 \wedge t_2) - \int_0^{s_1\wedge s_2} F(h(u); t_1)F(h(u); t_2)\, du. \tag{5.2}$$

For $g \in \Gamma$, define a two-parameter Gaussian process $W(s, t) = W(g; s, t)$ by specifying mean 0 and covariance Kg. If B_{st} is the usual Brownian sheet on the unit square (i.e., the continuous Gaussian process with mean 0, covariance $(s_1 \wedge s_2)(t_1 \wedge t_2)$) then using stochastic integrals for B_{st} (cf., Cairoli-Walsh, 1975) W has the representation

$$W(s,t) = \int_0^1 \int_0^1 I_{(0,s)}(u)\{I_{(0,F(g(u);t))}(v) - F(g(u); t)\}\, dB_{uv}.$$

From this representation a number of structural and sample function properties of this process are then immediate, but we do not dwell on these interesting probabilistic questions. The covariance of W shows this process to be closely related to the Kiefer process (Kiefer, 1972); indeed, the covariance of a Kiefer process is Kg with $g \equiv 1$ and F_θ the uniform distribution on $(0, \theta)$.

(5.3) PROPOSITION. (a) *Assume $F(\theta;\ \cdot)$ continuous. Under the measures $F(n;\ T(n, g);\ dx)$, $W_n(g;\ \cdot)$ converges weakly to $W(g;\ \cdot)$ on $C([0, 1] \times R)$.* (b) *There exist constants a and b such that whatever be g and $\{F(\theta;\ \cdot)\}$*

$$P_n\{|W_n|_m > t\} \leq ae^{-bt^2}$$

where P_n is the product measure of (a).

REMARKS. Part (a) was proved by Bickel and Wichura (1971) for the case g identically constant; the present more general case (independent but not identically distributed random variables) can be established by arguments within the framework of that paper. LeCam has shown me that an abstract variant of his Poissonization method will also prove (a). For the applications of this paper, one needs weak convergence only for the metric space L_m (instead of C); to prove this weaker result, the methods given in Millar (1981a) may be used. Finally, for the applications of Section 9, it is easy to see that the aforementioned proof establishes convergence if the distribution of X_{ni} is replaced by G_{ni}, where $|F(c(n;\ i;\ g)) - G_{ni}|_m \leq cn^{-1/2}$ and the product measure is the product of the G_{ni}. Part (b) of the theorem is familiar for the ordinary sample cdf from the work of Kiefer and Wolfowitz (1958). The present result is a special case of a remarkable recent result of LeCam, proved by an ingenious and elegant extension of his Poissonization technique to abstract valued random variables. According to LeCam, the norm in (b) may be replaced by the supremum norm, and the result holds with *no* hypotheses on the distributions of the X_{ni} other than independence. LeCam's proof will appear in his forthcoming monograph.

Proposition (5.3) immediately implies consistency results for minimum distance estimators. Let (Y, d) be a separable metric space with metric d, and for each n let $\{P_y^n : y \in Y\}$ be a family of probabilities on some measure space $(S^n, \mathscr{S}^n)$. A sequence of Y-valued random variables $\hat{Y}_n$ on S^n is called a $\sqrt{n}$ consistent estimator of $y \in Y$ if for every $\varepsilon > 0$ there exists $t > 0$ and n_ε such that

$$P_y^n\{n^{1/2}d(\hat{Y}_n, y) > t\} \leq \varepsilon \quad \text{for all} \quad y \in Y, \qquad n \geq n_\varepsilon.$$

There are many variants of this definition, most of them weaker by not requiring the uniformity over the entire parameter space. In classical parametric estimation theory, the existence of such estimators is often the first step to the development of efficient estimators. In our regression problem, the effective parameter set changes with n; it is $\{T(n, g), g \in \Gamma\}$ at time n. Accordingly, one naturally defines estimators $\hat{Y}_n$, with values in $\{T(n, g): g \in \Gamma\}$ to be $\sqrt{n}$ consistent relative to d, some metric if

$$\text{(5.4)} \qquad \lim_{t\uparrow\infty} \lim_{n\to\infty} \sup_{g\in\Gamma} P_{ng}\{n^{1/2}d(\hat{Y}_n, T(n, g)) > t\} = 0.$$

Here P_{ng} is the product measure (2.3). If the reparametrization of Section 3 is used, then the definition is changed in the obvious way: $\hat{Y}_n$ should take values in $\{FT(n, g);\ g \in \Gamma\}$, d should be a metric on L_m and $T(n, g)$ should be replaced by $FT(n, g)$ in (5.4).

Whether or not there are $\sqrt{n}$ consistent estimators depends strongly on the metric. With this in mind, the following proposition is of some interest.

(5.5) PROPOSITION. *Suppose Γ is dense in L. If the parameter set is $\{T(n, g), g \in \Gamma\}$ and if the metric there is $|\cdot|_L$ then there are no $\sqrt{n}$ consistent estimators.*

This may be proved with simple considerations involving the Hájek-LeCam asymptotic minimax theorem (Hájek, 1972; LeCam, 1972). From the point of view of this paper, $|\cdot|_L$ is in general the wrong metric to use on Γ. If this metric is weakened to that of convergence in measure, however, then it is often possible to find $\sqrt{n}$ consistent estimators. This fact follows from the next corollary, immediate from (5.3) but which also has a fairly simple direct proof. The minimum distance estimators in question are constructed from systems D_n with $D_n = \Gamma$ all n.

(5.6) PROPOSITION. *$\pi_n^0 \hat{F}_n$ is a $\sqrt{n}$ consistent estimator of the parameters $\{FT(n, g), g \in \Gamma\}$, whatever be Γ, if the metric is $|\cdot|_m$. If $F(\theta; t)$ is continuous in θ, then $\hat{g}_n$ (defined in Section 4) is $\sqrt{n}$ consistent estimator of the parameters $\{T(n, g); g \in \Gamma\}$, for the metric d_F.*

That the metric d_F is often equivalent to convergence in measure was discussed in Section 3. In short, it is impossible to find $\sqrt{n}$ consistent estimators of $T(n, g)$ if the metric is $|\cdot|_L$; but if we go to the "next" weakest metric, such estimators can be found.

6. Local structure and asymptotic distribution. This section shows that the minimum distance estimators $\pi_n^0\hat{F}_n$ defined in Section 4 converge, when suitably normalized, to certain functionals of the Gaussian process W. Since $\pi_n^0\hat{F}_n$ is L_m valued, the limiting distribution will, in general, be infinite dimensional.

We begin with the general formulation in Proposition (6.9). Because of the infinite dimensional character of these results, the hypotheses appear at first sight a bit complicated. That they can be verified in a number of interesting situations is illustrated in Sections 7 and 11. The case of finite dimensional Γ is particularly simple, and is singled out in Proposition (6.10). The basic results (6.9), (6.10) are asymptotic normality results for estimators in the reparameterized problem (parameter set $\{FT(n, g), g \in \Gamma\}$). These results can be rewritten to assert asymptotic normality of the minimum distance estimators of the orginal quantities $\{T(n, g), g \in \Gamma\}$; this is done in (6.11), (6.12), (6.15).

For the main result assume first that

(6.1) $$F_1(\theta; t) \equiv \partial/\partial\theta\, F(\theta; t) \quad \text{exists for each} \quad \theta \in \Theta.$$

Define, if possible, for each g with range in Θ, a mapping $A(g; \cdot)$ from L to R^2:

$$A(g; h)(s, t) = \int_0^s F_1(g(u); t)h(u)\, du.$$

Assume next that for each such g

(6.2) $$h \to A(g; h) \quad \text{is a bounded linear transformation from } L \text{ to } L_m.$$

This hypothesis requires $u \to F_1(g(u); t)$ to have mild integrability properties, as discussed in more detail in Section 11. The next assumption asserts, in particular, the Frêchet differentiability of the maps $g \to Fg$ of L to L_m:

(6.3) ASSUMPTION. Assume there exists an increasing function r on $(0, \infty)$ such that if $r_1(t) = t^{-1}r(t)$, then $r_1 \downarrow 0$ as $t \downarrow 0$ and for h, g

$$|Fh - Fg - A(g; h - g)|_m \le r(|g - h|_L).$$

Section 11 shows that in many of the cases of greatest practical interest, $r(t) \le t^{3/2}$. Variants of (6.3) more local in character and convenient for the study of exponential families are discussed in Section 11. Hypotheses (6.4)–(6.7) to follow should be skipped on first reading; basically they assert smoothness in both variables of $A(g; h)$ and they insist that the system $\{D_n\}$ not become complicated too fast. Assume for fixed $g_0 \in \Gamma$ that

(6.4) there exists a decreasing sequence $y_n > 0$ such that for any $g \in D_n\, |A(T(n, k_n, g_0); T(n, k_n, g - g_0))|_m \ge y_n |T(n; k_n; g - g_0)|_L$;

(6.5) there exists a decreasing sequence $z_n > 0$ such that $|FT(n; k_n; g) - FT(n; k_n; g_0)|_m \ge z_n r_1^{-1}(\frac{1}{2}y_n)$ for all $g \in D_n$ with $|T(n; k_n; g - g_0)|_L \ge r_1^{-1}(\frac{1}{2}y_n)$;

(6.6) the numbers z_n, y_n are subject to (a) $\lim n^{1/2}z_nr_1^{-1}(\frac{1}{2}y_n) = \infty$, (b) $\lim n^{1/2}r(cn^{-1/2}\cdot y_n^{-1}) = 0$ for all $c > 0$;

(6.7) (a) for every $c > 0$, $\sup_{g \in A(n, c)} |A(T(n; k_n; g_0); T(n; k_n; g)) - A(g_0; g)|_m \to 0$, where $A(n, c) = \{g \in D_n: |g - g_0|_L \le cy_n^{-1}\}$; (b) $|A(g_0; h - g_0)|_m \ge y_n |h - g_0|_L$ for $h \in D_n$;

(6.8) D_n is open in span D_n for the relative topology of L.

Define

$$\pi = \textit{orthogonal projection in } L_m \textit{ to the subspace } \{A(g_0; h): h \in \textit{span}\,\Gamma\} \equiv \tilde{\Gamma}.$$

That is, πx is the point of $\tilde{\Gamma}$ closest to x in the Hilbertian distance $|\cdot|_m$.

(6.9) PROPOSITION. *Fix $g_0 \in \Gamma$. Assume* (6.1)–(6.8). *Under the product measures $F(n; T(n, k_n, g_0); dx)$ and with respect to the metric $|\cdot|_m$:*

$$\pi_n^0 \hat{F}_n = FT(n; k_n; g_0) + \pi(\hat{F}_n - FT(n; k_n; g_0)) + o_p(n^{-1/2}).$$

In particular,

$$n^{1/2}(\pi_n^0 \hat{F}_n - FT(n; k_n; g_0)) \to \pi W$$

in distribution on L_m, where $W = W(g_0)$ is the two parameter Gaussian process defined in Section 5.

The proof is given in Section 10. As mentioned before, the assumption that Γ be finite dimensional results in substantial simplification.

(6.10) PROPOSITION. *Assume* (6.1)–(6.3), (6.7), *and that* $|F_1(\theta; t)| > 0$ *for* a.e. (θ, t). *Assume Γ finite dimensional and an open subset of span Γ. Assume convergence in the metric d_F on* sp Γ *implies convergence in measure. Take $D_n = \Gamma$ all n. Then the conclusion of* (6.9) *holds, with $k_n = n$.*

In the shift model, $F_1(\theta; t) = -f(t - \theta)$ where f is the density of F with respect to Lebesgue measure. According to Section 11, (6.1)–(6.3), (6.7) will hold if $\sup_t |f(t)| < \infty$, $\sup_t |f'(t)| < \infty$, and even weaker conditions suffice, depending on m. In the situation of (6.10), y_n may be taken constant in (6.4)–(6.7); the result itself follows from (6.9), as shown in Section 11.

We turn now to translating the foregoing asymptotic normality results back to the original parameter set $\{T(n, k_n, g): g \in \Gamma\}$. To simplify notation, take $k_n = n$. Consider first the simplest case: $\Gamma = \{\theta g_0, \theta \in I\}$ where I is a subinterval of the line and g_0 is a fixed element of L; this is the typical one dimensional Γ. Then asymptotically, if $\theta_0 g_0$ is fixed, $\pi_n^0 \hat{F}_n - FT(n, \theta_0 g_0)$ is projection onto the one dimensional subspace

$$\left\{\int_0^s F_1(\theta_0 g_0; t)(\theta - \theta_0) g_0, \theta \in R^1\right\},$$

so selection of $FT(n, \theta g_0)$ by $\pi_n^0 \hat{F}_n$ is a selection of $\theta - \theta_0$ in the evident way. By the usual characterization of projection in Hilbert space, if $\hat{\theta}_n$ is the θ "chosen" by $\pi_n^0 \hat{F}_n$, it is easy to see that $\hat{\theta}_n$ is asymptotically normal:

(6.11) COROLLARY. *Assume the hypotheses of* (6.10), *and that* $\Gamma = \{\Gamma g_0\}$. *Then under the product measures $FT(n, \theta_0 g_0)$*

$$\hat{\theta}_n - \theta_0 = \langle \hat{F}_n - FT(n, \theta_0 g_0), p \rangle_m / |p|_m^2 + o(n^{-1/2}),$$

where $p = A(\theta_0 g_0; g_0)$.

Evidently a similar argument works if Γ is k-dimensional. For later use, we record the result for $k = 2$. Suppose $\Gamma = \{\theta_1 g_1 + \theta_2 g_2: (\theta_1, \theta_2) \in I\}$ where I is an open convex subset of R^2 and $g_i \in L$ are fixed. Let $(\hat{\theta}_{n1}, \hat{\theta}_{n2})$ be the point chosen by $\pi_n^0 \hat{F}_n$. Then under the product measures associated with the point $\theta_0 g_1 + \theta_{00} g_2$

$$\begin{aligned} a(\hat{\theta}_{n1} - \theta_0) &= \langle \zeta_n, p_1 \rangle_m |p_2|_m^2 + \langle \zeta_n, p_2 \rangle_m \langle p_1, p_2 \rangle_m + o(n^{-1/2}), \\ a(\hat{\theta}_{n2} - \theta_{00}) &= \langle \zeta_n, p_1 \rangle_m \langle p_1, p_2 \rangle_m + \langle \zeta_n, p_2 \rangle_m |p_1|_m^2 + o(n^{-1/2}), \end{aligned} \tag{6.12}$$

where

$$\zeta_n = \hat{F}_n - FT(n, \theta_0 g_1 + \theta_{00} g_2), \quad p_\iota = A(\theta_0 g_1 + \theta_{00} g_2; g_\iota), \quad a = |p_1|_m^2 |p_2|_m^2 - \langle p_1, p_2 \rangle_m^2.$$

We conclude this section with one of the possible results asserting the asymptotic normality of $\hat{g}_n$, defined in Section 4. Assume $k_n = n$, and that the hypotheses of (6.9) hold. Assume also

$$\sup |A(T(n, g_0); T(n, h)) - A(g_0; T(n, h))|_m \to 0, \tag{6.13}$$

where the sup is over $h \in \Gamma$ with $|T(n, h)| \leq y_n^{-1}$; this condition is closely related to (6.7) and can be checked in the same way (Section 11). For convenience, assume $0 \in \Theta$, and that m puts a unit mass at $\{0\}$; this assumption can be removed at the price of a more extensive, but no more enlightening, argument. Assume

$$|F_1(\theta; 0)| > 0 \quad \text{all} \quad \theta. \tag{6.14}$$

Define a norm $|\cdot|_0$ on L by

$$|h|_0 = |A(g_0; h)(s, 0)|_L. \tag{6.15}$$

Because of (6.14), this indeed defines a norm, under which $L = \{h : \int h^2(u)\, du < \infty\}$ becomes a separable normed space but is not Banach. However $(L, |\cdot|_0)$ can be embedded isometrically and isomorphically in its second dual in the usual way; with this identification the closure of L—say $\bar{L}$—becomes a separable Banach space. Any L valued random variable can, by the aforementioned identification, be regarded as an $\bar{L}$ valued random variable. The random variable $\hat{g}_n$ in particular will be regarded this way.

(6.16) PROPOSITION. *With these understandings, $n^{1/2}(\hat{g}_n - g_0)$ converges under $F(n; T(n, g_0))$ to a normal distribution on $\bar{L}$.*

Actually this normal distribution is on span Γ, when Γ is regarded as a subset of $\bar{L}$. Notice also that if Γ is finite dimensional then the norms $|\cdot|_0$, $|\cdot|_L$ restricted to Γ, are equivalent.

This proposition is proved in Section 10; it is easy to calculate its characteristic functional using the facts in Sections 8 and 10.

7. Examples.

This section illustrates the application of the results of Section 6 to the two models of greatest practical interest—the shift model and the exponential family model. For a given Γ, there are in general an uncountable number of possible systems $\{D_n\}$ that will produce asymptotically normal estimators; the examples below illustrate only a very few of the possible choices.

7.A *Shift model.* Fix a distribution function F; for this model $F(\theta; t) = F(t - \theta)$. Assume F has a density f which is strictly positive on R^1, bounded, and has a bounded derivative f'. Then for any $g_0 \in L$, $A(g_0; h)(s, t) = \int_0^s f(t - g_0(u))h(u)\, du$ is a bounded linear operator, and differentiability hypothesis (6.3) holds with $r(t) =$ const. $t^{3/2}$; see Section 11 for details. This set of hypotheses is adopted for convenience only; one can weaken them at the price, for example, of putting hypotheses on m and/or adjusting the system $\{D_n\}$ so that its elements keep away from the infinities of f, f'. Assume the above for 7.A(i)–7.A(iv).

EXAMPLE 7.A(i). Suppose $\Gamma = \{\theta g_0 : \theta \in R^1\}$ where g_0 is a fixed element of L. If $g_0(s) = s$, this is the familiar straight line regression problem. If $g_0(s) = 1$, it is the familiar location model; see Millar, 1979b, for analysis of this particular case, together with discussion of the effect of various choices of the measure m. It is a simple matter to see that here all the hypotheses of (6.10) are satisfied.

EXAMPLE 7.A(ii). If $\Gamma = \{ag_1 + bg_2; g_i \in L; a, b \text{ real}\}$, then with g_i fixed the hypotheses of (6.10) again hold, and so the minimum distance estimator of (a, b) is asymptotically normal, with asymptotic expansion given by (6.12). The standard straight line regression has $g_1 = 1$, $g_2(s) = s$, and, unlike the least squares estimate, the minimum distance estimate will be robust (cf. Section 10).

EXAMPLE 7.A(iii). Let Γ consist of all $g \in L$ which are bounded and have a bounded derivative g'. Notice that Γ is dense in L. Let F be $N(0, 1)$. At time n, the possible distributions of the data are by hypothesis to be given by product measures $F(n;\ T(n;\ k_n;\ g))$, $g \in \Gamma$, where $k_n = n^{1/6}$. Let D_n consist of all elements $g \in \Gamma$ with $|g|_K \leq (2 \log \log n)^{1/2}$, $|g'|_K \leq M_n$ where M_n is any fixed sequence such that $M_n n^{-1/6} \to 0$ and where $|h|_K = \sup_t |h(t)|$. Since $r(t) = \text{const.}\ t^{3/2}$ in this case, it is not difficult to use the suggestions of Section 11 to check that all the hypotheses of (6.9) are satisfied. Accordingly, the minimum distance estimate of $\{FT(n;\ k_n;\ g) : g \in \Gamma\}$ is asymptotically normal, and will be robust according to the definition of Section 10; the normal distribution in question is, of course, infinite dimensional.

EXAMPLE 7.A(iv). Let F again be $N(0, 1)$. Let $\mathcal{F}_n$ be the sigma field on $[0, 1]$ generated by intervals $(i/n, i + 1/n]$. Take Γ to consist of all g such that g is $\mathcal{F}_{2^k}$-measurable for some k. Notice again that Γ is dense in L. At stage n, the distribution of the data is given by the product measures $F(n;\ T(n, g))$ for some $g \in \Gamma$. If D_n consists, for example, of all $\mathcal{F}_{a_n}$-measurable functions bounded by $(2 \log \log n)^{1/2}$, where $a_n = \exp(k_n \log 2)$, $k_n = (\log n)/6 \log 2$ then the assumptions of (6.9) hold.

7.B *Exponential model.*

EXAMPLE 7.B(i). Let $F(\theta;\ t)$ be the cdf of the exponential family density $\exp\{\theta q(u) + b(\theta)\}$, where $\Theta = (a, b)$. The one dimensional Γ has the form $\Gamma = \{\theta g_0,\ \theta \in I\}$ for some fixed $g_0 \in L$ having range in (a, b). Here I is to be an interval determined by

$$\{\theta\colon a < \inf_u \theta g_0(u) \leq \sup_u \theta g_0(u) < b\}.$$

It is possible to allow, for example, $g_0(u) > a$ but $\inf_u g_0(u) = a$; we shall not discuss this case here. The operator $A(g;\ h)$ is given formally by

$$\int_a^t \int_0^s h(u)\{q(x) + b'(g(u))\}\exp\{g(u)q(x) + b(g(u))\};$$

if $F(\theta;\ t) = 1 - e^{-\theta t}$ with $\Theta = (0, \infty)$, then

$$A(g;\ h)(s, t) = \int_0^s -te^{-g(u)t}\, du.$$

If m is arbitrary but finite, and if only range $g \subset \Theta$, this operator need not be bounded. For Γ described above, and with $g = \theta_0 g_0$ for some $\theta_0 \in I$, it is easy to see that it is bounded. The hypotheses of (6.10) are then easily checked in this case and the estimate of θ is asymptotically normal with distribution given by (6.11). The general finite dimensional case can be treated in the same way, if I is replaced by the evident analogue.

EXAMPLE 7.B(ii). Suppose Γ consists, as in 7.A(iii), of all $g \in L$ which are bounded and have a bounded first derivative g'. For simplicity let us begin with the model $F(\theta;\ t) = 1 - \exp(-\theta t)$ which already exhibits all the features of the general case. The model at stage n will be given by the product measures $F(n;\ T(n,\ k_n;\ g))$, $g \in \Gamma$ and where $k_n = n^{1/6}(\log n)^{-3}$. To find a suitable system D_n, we work with the variant of Proposition (6.9) given in Section 11. If $\Theta_n = [a_n, \infty)$ and we arbitrarily choose $a_n = (\log n)^{-1}$ then working through Section 11 and making several other more or less arbitrary choices, one is led to the following choice of D_n as one of the many possibilities:

$$D_n = \{g \in L : |g'|_k \leq n^{1/6}(\log n)^{-4},\ (\log n)^{-1} \leq g \leq \log \log n\}.$$

With D_n so chosen, the hypotheses of the modified (6.9) are satisfied and so the minimum distance estimator based on D_n will be asymptotically normal and robust. Evidently one can alter the upper and lower bounds in D_n quite a bit and still get another system that works.

For the general exponential model with $\Theta = (a, b)$ one again chooses subintervals $\Theta_n = (a_n, b_n)$, $a_n \downarrow a$, $b_n \uparrow b$ such that on Θ_n both F_1, F_2 are bounded, say, by c_n; here $F_2(\theta; t) = \partial F_1(\theta; t)/\partial\theta$. Using standard properties of exponential families (cf. Johansen, 1979) one easily sees that this can be done, and that a_n, b_n can be chosen so that c_n increases as slowly as desired. Moreover the function r associated with each Θ_n as in Section 11.A will be $r(t) = t^{3/2}$. It is then routine to arrange (11.4a), (11.4b) and to ensure the other hypotheses. In practice, one would no doubt replace each D_n just constructed by a finite dimensional subset D'_n thereof, such that still $D'_n \uparrow \Gamma$.

EXAMPLE 7.B(iii). Bring in again the model $F(\theta; t) = 1 - \exp(-\theta t)$ and as in 7.A(iv) take Γ to be all $g \in L$ such that g is $\mathscr{F}_{2^n}$-measurable for some n. The distribution of the data is given by the product measures $F(n; T(n, g))$, $g \in \Gamma$. If D_n consists, for example, of all $g \in L$ such that $(\log n)^{-1} \leq g \leq \log\log n$ and g is $\mathscr{F}_{a_n}$-measurable, where $a_n = \exp(c_n \log 2)$, $c_n = (\log n)/(7 \log 2)$, then the hypotheses of (6.9) will be satisfied, and so the minimum distance estimator is robust and asymptotically normal.

8. Gaussian experiments indexed by Hilbert space. This section introduces statistical experiments parametrized by the points of a certain Hilbert space H_μ. Such statistical experiments are key prerequisites for the study of robustness given in Section 9. As a by-product, this section furnishes added structural information about the Gaussian process $W(g; s, t)$ of Section 5. The function $g_0 \in \Gamma$ will be fixed throughout this section and the next; it will not always appear explicitly in notations which actually depend on it.

To begin, suppose that each $F(\theta; t)$ has a density $f(\theta; t)$ with respect to a sigma finite measure μ on the line such that $f(\theta; t) > 0$ a.e. μ. Define H_{00} to be the subspace of $L^2([0, 1] \times R, ds\, d\mu)$ consisting of all finite linear combinations of elements in L^2 of the form $e(u)h(g_0(u); v)$ where e is a function on $[0, 1]$, h is a function on $[0, 1] \times R$ and

$$\int h(g_0(u); v) f^{1/2}(g_0(u); v)\mu(dv) = 0$$

for almost all u. Define

$$H_\mu = \text{closure of } H_{00} \text{ in } L^2(ds\, d\mu). \tag{8.1}$$

Then H_μ is a Hilbert space; denote by $\langle , \rangle_\mu$ its inner product. Define the bounded one-to-one linear operator M mapping H_μ to L^2 by

$$(Mh)(s, t) = \int_0^s \int_0^t f^{1/2}(g_0(u), v)h(u, v)\, du\mu(dv). \tag{8.2}$$

Of course, M depends on g_0. This operator M is closely related to the operator $A(g_0; h)$ on L. Indeed, define $\dot{f}(\theta; t) = \partial f(\theta; u)/\partial\theta$, and for $h \in L$,

$$h_f(u, v) = \dot{f}(g_0(u); v)[f(g_0(u); v)]^{-1/2}h(u).$$

Assuming the evident integrability properties, $h_f \in H_\mu$ and $A(g_0; h) = Mh_f$.

Define for each s, t the function

$$h_{st}(u, v) = I_{(0,s)}(u) f^{1/2}(g_0(u); v)[I_{(-\infty,t]}(v) - F(g_0(u); t)]. \tag{8.3}$$

Then each $h_{s,t}$ belongs to H_μ and by direct calculation

$$Mh(s, t) = \langle h_{st}, h\rangle_\mu, \qquad h \in H_\mu \tag{8.4}$$

and, if $K(g_0; s_1, t_1; s_2, t_2)$ is the covariance of the Gaussian process Wg_0 of Section 5, then again by direct calculation

$$(8.5)\qquad K(g_0; s_1, t_1; s_2, t_2) = \langle h_{s_1t_1}, h_{s_2t_2}\rangle_\mu .$$

Let Q be the standard normal cylinder measure on H_μ; that is, Q is a finitely additive measure on the cylinder sets of H_μ with characteristic functional $\varphi(h) = \exp\{-\tfrac{1}{2}|h|_\mu^2\}$, $h \in H_\mu$. Let Q_0 be the image of Q under the mapping M. On the basis of the preceding paragraph, it is easy to see that

(8.6) $\qquad Q_0$ is the distribution of Wg_0.

Define for h in H_μ, probabilities Q_h on the Borel sets A of L^2 by

$$(8.7)\qquad Q_h(A) = Q_0(A - Mh).$$

Then by basic theory the measures Q_h are countably additive, mutually absolutely continuous, and the distribution of $\log(dQ_h/dQ_0)$ under Q_0 is $N(-\tfrac{1}{2}|h|_\mu^2, |h|_\mu^2)$. See Millar (1979a) and references therein for discussion of these notions.

Next, for each $h \in H_{00}$ and each n, we construct probabilities Q_h^n such that $\lim_n dQ_h^n/dQ_0^n = dQ_h/dQ_0$ in distribution, under $\{Q_0^n\}$. These measures figure in the robustness analysis of Section 9. Suppose $h(u, v) = \sum a_{ij}e_i(u)h_j(g_0(u), v)$, a *finite* sum; this is the typical element of H_{00}. For such an h, construct densities $f_{n,i,h}$ by the recipe

$$(8.8)\qquad f_{n,\imath,h}^{1/2}(x_i) = (1 - |q_{nih}|_\mu^2)^{1/2}f^{1/2}(c(n; i; g_0); x_i) + q_{n\imath h}(x_i),$$

where $c(n; i; g_0)$ was defined in (2.2) and

$$q_{nih}(x_i) = \frac{1}{2}\textstyle\sum_{j,k} a_{kj}b(n; i; k)h_j(c(n; i; g_0); x_i),$$

$$b(n; i; k) = n^{1/2}\int I_{(\imath-1)/n,i/n}(u)e_k(u)\,du.$$

Define

$$(8.9)\qquad Q_h^n = \text{product measure on } R^n \text{ having density } \textstyle\prod_i f_{nih}(x_i).$$

(8.10) PROPOSITION. *Under the measures* $\{Q_0^n\}$, $\log(dQ_n^n/dQ_h^n)$ *is asymptotically normal* $N(-\tfrac{1}{2}|h|_\mu^2, |h|_\mu^2)$.

PROOF. First one shows $\sup_i|q_{nih}|_\mu \to 0$ as $n \to \infty$. To see this, it suffices to show (since range of k, j is finite) that for each k $\sup_i|b(n; i; k)| \to 0$ as $n \to \infty$. This last is immediate if e_k is bounded; for general e_k it may be proved by approximating e_k in norm by bounded functions. Define $Y_{nih} = (f_{nih}/f_{ni0})^{1/2}(x_i) - 1$. From the estimate just made,

$$\textstyle\sum_i Y_{nih} = \sum_{i=1}^n q_{nih}(x_i)f_{ni0}^{-1/2}(x_i) - \frac{1}{2}\sum|q_{nih}|_\mu^2 + o(1).$$

On the other hand, straightforward computation shows that

$$\lim_{n\to\infty}\textstyle\sum_i|q_{nih}|_\mu^2 = \frac{1}{4}|h|_\mu^2.$$

This computation may be completed by multiplying out everything, and treating the sum on i before those on k, j; the computation is facilitated by using elementary properties of the conditional expectations $E(\cdot\,|\,\mathscr{F}_n)$ of (2.4). From the last two displays and the central limit theorem, $\sum_i Y_{nih}$ is asymptotically $N(-(1/8)|h|_\mu^2, \tfrac{1}{4}|h|_\mu^2)$. The proof may be completed using the development of LeCam (1969).

9. Robustness. This section explains the precise sense in which the minimum distance estimators of this paper are robust. Assume for convenience that $k_n = n$.

If $g_0 \in \Gamma$ is "true," the formulation of Section 2 alleges that the observations at time n follow the product measure $\prod F(c(n, i, g_0), dx)$. It is a familiar fact in practical statistics that, due to "data contamination," the observations may not follow this distribution precisely for any choice of $c(n; i; g)$. Causes of such data contamination can be attributed to many factors, including human error, roundoff error, and so forth. An estimator will be robust if its performance does not deteriorate when the assumptions of the model are not satisfied precisely. There are many ways of giving technical formulation to the various possibilities. Following Beran (1981), Millar (1981a), Rieder (1981), we will give robustness a purely decision theoretic formulation, where optimality ($\equiv$ robustness) is to be a local asymptotic minimax property. The advantages of such an approach are discussed at great length in Millar (1981a); since the formulation here is similar to that of the earlier paper, the discussion will be brief.

To start, fix $g_0 \in \Gamma$. For each n and $i \le n$, bring in functions $S(n, i)$ on R^1, and define

$$F_n(S(n, i); t) = F(c(n; i; g_0); t) + n^{-1/2} S(n, i)(t). \tag{9.1}$$

Assume $S(n, i)$ chosen so this definition gives a bona fide probability distribution function. Definition (9.1) gives the possible perturbations to the distribution of the ith observation that can be attributed to "data contamination." The distribution of the data actually observed then follows $F^n(S(n, \cdot); dx) \equiv \prod_i F_n(S(n, i); dx_i)$, $x = (x_1, \cdots, x_n)$. The assumption of independence can be relaxed, as pointed out in Section 13. The distributions of Section 8 having density of the ith observation given by f_{nih} have this form with $S(n, i)(t)$ given essentially by $2n^{1/2} \int^t f^{1/2}(c(n, i, g_0); u) q_{nih}(u)\, du$. The formulation of "data contamination" here includes the possibility that either the regression function g_0 is subject to error, or the parametric family $F(\theta; \cdot)$ is inexact, or, of course, both.

The possible distributions $F^n(S)$ of the data at time n having been specified, the task is still to estimate the functions $F(T(n, g))$, as described in Section 3. Given cdf's $F_n(S(n, i); t)$ as in (9.1) define

$$\bar{F}(n; S; t) = n^{-1} \textstyle\sum_{i \le ns} F_n(S(n, i); t) \tag{9.2}$$

a cdf on $[0, 1] \times R$. If ℓ is a nondecreasing function on $[0, \infty)$, then a reasonable measure of loss, when $F^n(S)$ is the distribution of the data given by (9.1) and $FT(n, g)$ is chosen from the decision space is

$$\ell(n^{1/2} | FT(n, g) - \bar{F}(n, S) |_m).$$

The $n^{1/2}$ in this recipe is reasonable, because according to Section 5, $\pi_n^0 F_n$ can estimate the parameters $FT(n, g)$ this closely. It is easy to see, however, that (just as for the location problem) the present problem is degenerate for this precise loss function (the minimax risk is easily seen to be $\ell(\infty)$), so we modify it in the manner of Millar (1981a); many other loss functions are possible. Define loss at stage n to be

$$\ell(n^{1/2} | \pi_n^0 \bar{F}(n, S) - FT(n, g) |_m) = \ell_n(S, FT(n, g)). \tag{9.3}$$

The reasonable character of similar loss functions is defended in Millar (1981a) on the grounds that they can penalize both for bad data and for an inept estimate. With the foregoing decision theoretic set-up, call an estimator T_n^0 with values in $\{FT(n, g), g \in \Gamma\}$ robust if it is locally asymptotically minimax:

$$\begin{aligned} &\lim_{c\uparrow\infty} \lim_{n\to\infty} \inf_{\{T_n\}} \sup_{S: |S|_m \le c} \int \ell_n(S, T_n)\, dF^n(S(n, \ ), \cdot). \\ &= \lim_{c\uparrow\infty} \lim_{n\to\infty} \sup_{S: |S|_m \le c} \int \ell_n(S, T_n^0)\, dF^n(S(n, \ ), \cdot). \end{aligned} \tag{9.4}$$

Here the infimum in the first expression is over all estimates of the parameters $\{FT(n, g), g \in \Gamma\}$.

That this definition of robustness captures the basic stability concept desired has been argued at length by Millar (1981a), Beran (1981). Evidence that it works in the present context is provided by the following proposition.

(9.5) Proposition. *Let $F(\theta, dt)$ be the shift model; suppose the assumptions of* (6.9) *are satisfied. Let $\hat{\xi}_n$ be the least squares estimate of $\{T(n, g), g \in \Gamma\}$ so that $F\hat{\xi}_n$ is the estimator of the reparametrized problem described in Section* 3. *Then $F\hat{\xi}_n$ is not robust in the sense of Definition* (9.4).

This can be extended to exponential models, and so forth; see Millar (1981a), to see how it can be proved for the special case of location models. The counter examples themselves demonstrate that it is the outliers that disrupt the estimate $F\hat{\xi}_n$. The main result of this section is that the minimum distance estimators of Section 6 are robust in the sense of definition (9.4).

(9.6) Proposition. *Assume* (6.1)–(6.8); *assume as in Section* 8, *that each $F(\theta; t)$ has a density $f(\theta; t) > 0$ with respect to a measure μ. Assume further that*

$$A(g_0, h) \subset M(H_\mu) \textit{ for any } h \in \operatorname{span} \Gamma \tag{9.7}$$

and that the increasing function ℓ of (9.3) *satisfies the growth condition*

$$y^{-2} \log \ell(y) = o(1) \quad \text{as} \quad y \to \infty.$$

Then $\pi_n^0 F_n$ is robust in the sense of (9.4) *and the limiting minimax value is $E\,\ell(|\pi W g_0|_m)$.*

A trivial condition for (9.6) was suggested in Section 8; see Millar (1981a) for other possibilities.

Proof. (Sketch). It may be shown that the basic decision theoretic structure forces good estimators to choose elements of the decision space of the form $FT(n, g_0) + n^{-1/2}A(g_0; h) + o(n^{-1/2})$. With some effort, unenlightening in itself, it may then be shown that the loss (9.3) may be approximated asymptotically by the local loss

$$\ell(|\pi[A(g_0, g) - \Delta(n, S)]|_m) \quad \text{where} \quad \Delta(n, S) = n^{1/2}(\bar{F}(n, S) - FT(n, g_0)).$$

This reduction will use (5.3b). One may now calculate the value of the second expression in (9.4) using the remarks after (5.3) concerning weak convergence in L_m. The value of the first expression in (9.4) may be deduced, as in Millar (1979a, 1981a), from the Hájek-LeCam asymptotic minimax theorem, using the experiments $\{Q_h^n, h \in H_{00}\}$; for this it is necessary to note that $x \to \ell(|\pi x|_m)$ is subconvex on L_m.

Remark. Considerable work can be avoided if it is agreed that since the formulation is local, the analysis should therefore begin with the local loss function given in the proof (thus by-passing the complicated reduction beginning with a global loss function). The local decision space is then $\{A(g_0; g)\}$ and the (local) decision made by $\pi_n^0 \hat{F}_n$ is then $\pi W_n g_0$. This simplified formulation seems to entail no real conceptual loss.

10. Proofs of (6.9) and (6.15).

Proof of (6.9). The proof bears many points in common with other analyses of minimum distance estimators (cf. Beran, 1977; Millar, 1981a; Bolthausen, 1977; Pollard, 1980; Wolfowitz, 1957). A recent bibliography of Parr (1980) surveys current work on minimum distance estimators. For a one dimensional shift model, Williamson (1979), gives asymptotically normal estimates without complete specification of the distribution F, but no optimality.

Assume n so large that $g_0 \in D_n$. Let $a_n > 0$ and set $B_n = \{g \in D_n: |T(n; k_n; g - g_0)|_L \geq a_n\}$. By the triangle inequality, (6.5) and the choice of $a_n = r_1^{-1}(\tfrac{1}{2}y_n)$,

$$\inf_{g \in B_n} |\hat{F}_n - FT(n; k_n; g)|_m \geq z_n r_1^{-1}(\tfrac{1}{2}y_n) - n^{-1/2}|W_n|_m$$

where W_n was defined in Section 5. Because of (6.6a) and (5.3), this implies with probability approaching 1 as $n \to \infty$:

$$\inf_{g \in D_n} n^{1/2} |\hat{F}_n - FT(n; k_n; g)|_m = \inf_{g \in C_n} n^{1/2} |\hat{F}_n - FT(n; k_n; g)|_m$$

where $C_n = D_n - B_n$. In short, the minimum distance estimator must be computed only on g's "within" a_n of g_0. By (6.3) and the choice of a_n

$$|\hat{F}_n - FT(n; k_n; g)|_m \geq \tfrac{1}{2} y_n |T(n; k_n; g - g_0)|_L - n^{-1/2} |W_n|_m.$$

Accordingly, $\pi_n^0 \hat{F}_n$ must be computed by taking an infimum over only those g's $\in C_n$ such that $|T(n; k_n; g - g_0)|_L \leq 4n^{-1/2} |W_n|_m y_n^{-1}$. For such g's, relative to the metric $\|_m$,

$$FT(n; k_n; g) = FT(n; k_n; g_0) + A(T(n; k_n; g_0); T(n; k_n; g - g_0)) + o_p(n^{-1/2})$$

provided that $n^{1/2} r(cn^{-1/2} y_n^{-1}) \to 0$ for every c as $n \to \infty$; but hypothesis (6.6b) guarantees that. It follows that within $o_p(n^{-1/2})$

$$\pi_n^0 \hat{F}_n = FT(n; k_n; g_0) + n^{-1/2} A(T(n; k_n; g_0); T(n; k_n; \hat{h})) + o(n^{-1/2})$$

where $\hat{h} \in \operatorname{span} D_n, |T(n; k_n; \hat{h})|_L \leq 4 |W_n|_m y_n^{-1}$. Because of (6.7a)

$$\pi_n^0 \hat{F}_n = FT(n; k_n; g_0) + n^{-1/2} A(g_0; \hat{h}) + o(n^{-1/2}).$$

Next, let π_n be orthogonal projection in the Hilbert space $L^2([0, 1] \times R, dsdm)$ to the subspace $A_n = \{A(g_0; h): h \in \operatorname{span} D_n\}$ and π_n^1 orthogonal projection to $FT(n; k_n; g_0) + A_n$. Because of $\sqrt{n}$ consistency, $|\hat{F}_n - FT(n; k_n; g_0)|_m \leq |W_n|_m n^{-1/2}$, so by Pythagoras one can compute $\pi_n^1 \hat{F}_n$ by looking at those h such that $|A(g_0; h - g_0)|_m \leq |W_n|_m n^{-1/2}$. By (6.7b), one need examine only those h's such that $|h - g_0| \leq |W_n| n^{-1/2} y_n^{-1/2}$. It follows that

$$\begin{aligned}\pi_n^1 \hat{F}_n &= FT(n; k_n; g_0) + \pi_n(\hat{F}_n - FT(n; k_n; g_0)) \\ &= FT(n; k_n; g_0) + n^{-1/2} A(g_0; \underline{h}),\end{aligned}$$

where $\underline{h} \in D_n$, $|\underline{h}| \leq 4 |W_n|_m y_n^{-1}$. By (6.4a), (6.3), (6.8)

$$\begin{aligned}\pi_n^1 \hat{F}_n &= FT(n; k_n; g_0) + n^{-1/2} A(T(n; k_n; g_0); T(n; k_n; \underline{h})) + o(n^{-1/2}) \\ &= FT(n; k_n; g_0 + n^{-1/2} \underline{h}) + o(n^{-1/2}).\end{aligned}$$

Next, since π_n^1 is a minimum distance operation,

$$\begin{aligned}|\hat{F}_n - \pi_n^0 \hat{F}_n|_m &= |\hat{F}_n - FT(n; k_n; g_0) + A(g_0; \hat{h}) n^{-1/2}|_m + o(n^{-1/2}) \\ &\geq |\hat{F}_n - \pi_n^1 \hat{F}_n|_m + o(n^{-1/2})\end{aligned}$$

and, since π_n^0 is also a minimum distance operation,

$$\begin{aligned}|\hat{F}_n - \pi_n \hat{F}_n|_m &= |\hat{F}_n - FT(n; k_n; n^{-1/2} \underline{h} + g_0)|_m + o(n^{-1/2}) \\ &\geq |\hat{F}_n - \pi_n^0 \hat{F}_n|_m + o(n^{-1/2}).\end{aligned}$$

Therefore

$$|\hat{F}_n - \pi_n^1 \hat{F}_n|_m = |\hat{F}_n - \pi_n^0 \hat{F}_n|_m + o(n^{-1/2}).$$

By Pythagoras it then follows that

$$|\pi_n^0 \hat{F}_n - \pi_n^1 \hat{F}_n|_m^2 = |\hat{F}_n - \pi_n^0 \hat{F}_n|_m^2 - |\hat{F}_n - \pi_n^1 \hat{F}_n|_m^2 + o(n^{-1/2})$$

so

$$\pi_n^0 \hat{F}_n = FT(n; k_n; g_0) + \pi_n(\hat{F}_n - FT(n; k_n; g_0)) + o(n^{-1/2}).$$

Proposition 6.9 is then immediate from the following lemma, given without proof.

LEMMA. *Let G_n, $n = 1, 2, \cdots$, be closed convex sets in a separable Hilbert space H with norm $|\cdot|_H$ such that $G_n \subset G_{n+1}$. Let G be the closure of $\cup G_n$. Let π, π_n be projection to G, G_n respectively. Then* $\lim_n |\pi_n x - \pi x|_H = 0$, *uniformly for x in compact subsets of H.*

In our application of this lemma, the G_n are in fact subspaces; in this case the proof of the lemma is somewhat simpler.

PROOF OF (6.15). Define V to be the mapping of $B_0 = \{A(g_0; h): h \in L\}$ to L by $V: A(g_0; h) \to h$. Notice that B_0 is a subspace, not closed in general. Since $A(g_0; h)$ determines h, this mapping is well defined and linear. Since m has a point mass at 0, $|Vx|_0 \leq |x|_m$ for $x \in B_0$. It follows that V may be extended as continuous linear map from B, the closure of B_0, to $\bar{L}$. Because of (6.13),

$$n^{1/2}[\pi_n^0 \hat{F}_n - FT(n; g_0)] = A(g_0; n^{1/2}T(n; \hat{g}_n - g_0)) + o(1).$$

But $A(g_0; n^{1/2}T(n; \hat{g}_n - g_0))$ is asymptotically normal, according to (6.9). Since V is a bounded linear operator, $V(A(g_0; n^{1/2}T(n; \hat{g}_n - g_0)))$ is asymptotically normal; i.e., $n^{1/2}(\hat{g}_n - T(n; g_0))$ has an asymptotically normal distribution on $\bar{L}$.

Of course, if φ is the characteristic functional of Wg_0, then the ch. f. of this normal distribution is $\varphi(V^*h)$ where the asterisk denotes adjoint and where h is in the dual of $\bar{L}$.

11. Discussion of hypotheses (6.1)-(6.8).

This section gives easily verifiable criteria under which hypotheses (6.1)-(6.8) will be satisfied. There are many possible combinations: one can place hypotheses on the family $\{F(\theta; t)\}$ and can in a number of ways vary k_n, D_n as well. Accordingly, the matter is somewhat complicated, and the section is fairly technical.

11A. HYPOTHESIS (6.2). Whether or not (6.2) holds depends in general on F_1, g and m. By Schwarz, $|A(g; h)|_m \leq C(g, m)|h|_L$, where $C^2(g, m) = \int\int F_1^2(g(u); t)\, du\, m(dt)$, so a simple crude condition for (6.2) is $C(g, m) < \infty$.

One way to achieve this is to consider only g's such that $\sup_{\theta \in R(g)} |F_1^2(\theta; t)| \leq M(g)$ where $R(g)$ = range g, and $M(g)$ is a constant independent of t. Then it does not matter what m is. For example, in the shift model, $F(\theta; t) = F(t - \theta)$; if F has density f then $F_1(\theta; t) = -f(t - \theta)$ and so it suffices that f be bounded. In the model $F(\theta; t) = 1 - \exp(-\theta t)$, $C(g, m)$ will be finite for any m provided $R(g) \subset [\varepsilon, \infty)$ for some $\varepsilon > 0$.

On the other hand, one can achieve $C(g, m) < \infty$ for all g by insisting that m satisfy $\int \sup_\theta F_1^2(\theta; t) m(dt) < \infty$. For example, in the shift model just mentioned, it would be enough if m has a bounded density with respect to Lebesgue measure and $\int f^2(t)\, dt < \infty$. In the exponential model of the preceding paragraph $C(g, m) < \infty$ for all g if, for example, $\int t^2 m(dt) < \infty$.

11B. HYPOTHESIS (6.3), *differentiability*. The following criterion and its variant in 11B(a) below suffices to handle many cases of practical interest.

(11.1) PROPOSITION. *Assume $F_1(\theta; t)$ has, for each t, a modulus of continuity $w(t; u)$: $|F_1(\theta; t) - F_1(\theta'; t)| \leq w(t; |\theta - \theta'|)$. Assume further that if $w(u) = \int w(t; u) m(dt)$, then $\lim_{u \to 0} w(u) = 0$. Assume that $S(t) \equiv \sup_\Theta |F_1(\theta; t)| \in L_1(R^1, dm)$. Then* (6.3) *holds with $r(u) = u\{w(u^{1/2}) + Su^{1/2}\}$, where S is a constant multiple of $\int S(t) m(dt)$.*

(11.2) COROLLARY. *If $F_2(\theta; t) = \partial F_1(\theta; t)/\partial\theta$ and if $\sup_\Theta |F_2(\theta; t)|$, $\sup_\Theta |F_1(\theta; t)|$ are bounded independently of t, then* (6.3) *holds with $r(u)$ equal to a constant multiple of $u^{3/2}$.*

For example, in the shift model $F(\theta; t) = F(t - \theta)$, the condition of (11.2) will hold if F has a density f having a bounded derivative. Proof of (11.1) is given at the end of this subsection.

11B(a) A VARIANT. The following variant of hypothesis (6.3) is indispensible for the study of general exponential models. The altered assumption puts a condition on the sets D_n. *Fix $g_0 \in \Gamma$; assume there exists an increasing function r on* $[0, \infty)$ *such that if $r_1(t)$*

$= t^{-1}r(t)$, *then* $r_1 \downarrow 0$ *as* $t \downarrow 0$ *and for* $h \in D_n$ *there are numbers* p_n *such that*

$$(11.3) \qquad |Fh - Fg_0 - A(g_0, h - g_0)|_m \le p_n r(|h - g_0|_L).$$

Under (11.3), *Proposition* (6.9) *continues to hold provided* y_n, z_n *are subject to*

$$(11.4a) \qquad \lim n^{1/2} z_n r_1^{-1}(\tfrac{1}{2} y_n / p_n) = \infty,$$

$$(11.4b) \qquad \lim n^{1/2} p_n r(cn^{-1/2} y_n^{-1}) = 0 \quad \textit{for every} \quad c > 0.$$

To see this, essentially the same proof may be given, but with a_n there taken as $r_1^{-1}(\tfrac{1}{2} y_n/p_n)$.

This variant may be applied as follows. If Θ_n is a convex subset of Θ, $\Theta_n \uparrow \Theta$, suppose

$$\sup_{\theta \in \Theta_n} |F_2(\theta; t)| \le b_n, \ \sup_{\theta \in \Theta_n} |F_1(\theta; t)| \le b_n$$

for a sequence of real numbers b_n. Assume all $g \in D_n$ have range in Θ_n. Then a simple variant of (11.2) shows that (11.3) holds with $p_n = b_n$ and $r(t)$ a constant multiple of $t^{3/2}$. For example, in the model $F(\theta; t) = 1 - \exp(-\theta t)$, one could take $\Theta_n = [a_n, \infty)$ where a_n is a sequence of numbers decreasing to 0; then $b_n = (a_n)^{-2}$ suffices to bound F_1 and F_2 over Θ_n. Notice that by proper choice of a_n, the numbers b_n can be taken to increase to infinity as slowly as desired.

11B(b) Proof of (11.1). Fix s, t. Then

$$\begin{aligned} &|Fh(s, t) - Fg(s, t) - A(g; h - g)(s, t)| \\ &\qquad = \left| \int_0^s \int_0^{h(a)-g(a)} \{F_1(g(a) + u; t) - F_1(g(a); t)\} du\, da \right| \\ &\qquad \le |w(t; |h - g|)|_L |h - g|_L. \end{aligned}$$

If $\lambda > 0$, the Chebychev inequality implies

$$\begin{aligned} \int_0^1 w^2(t; |h(s) - g(s)|)\, ds &\le w^2(t; \lambda |g - h|_L) + 4C^2(t) P\{s : |h(s) - g(s)| > \lambda |h - g|_L\} \\ &\le w^2(t; \lambda |g - h|_L) + 4C^2(t)\lambda^{-2}. \end{aligned}$$

Choose $\lambda = |h - g|_L^{-1/2}$ to complete the verification.

11C. Hypotheses (6.4)–(6.6), $k_n = n$. Let C_n, $n = 1, 2, \cdots$, be a sequence of closed convex subsets of L, $C_n \subset C_{n+1}$. Suppose $\Gamma = \cup C_n$. This subsection shows under mild hypotheses that it is always possible to find integers $n_1 < n_2 < \cdots$ such that if $D_n = C_j$ when $n_j \le n < n_{j+1}$, then hypotheses (6.4)–(6.6) hold with $k_n = n$. That is, almost any increasing family of convex subsets D_n can be used, and (6.4)–(6.6) will still hold provided that D_n increases slowly enough.

To formulate the requisite hypotheses, assume (6.1)–(6.3) and define $S_n h = A(T(n; g_0); T(n; h))$, $Sh = A(g_0; h)$. Assume next that

(11.5) each C_j is finite dimensional,

and

(11.6) the bounded linear operators S_n, S are $1 - 1$ on span C_j, for each j and n sufficiently large.

Since C_j is finite dimensional, the condition $T(n; h_1) = T(n; h_2)$ will imply $h_1 = h_2$ for $h_i \in C_j$ provided n is chosen large enough; assume henceforth that n is so chosen. Then it is easy to see that (11.6) can be guaranteed if $F_1(g(s); t)h_1(s) = F_1(g(s); t)h_2(s)$ for all t, a.e. s implies $h_1 = h_2$ a.e. This happens, of course, if $|F_1(\theta; t)| > 0$ for a.e. θ, t. For example, in the shift model, (11.6) is guaranteed if F has a density f that is strictly positive.

Assume further that

$$\text{(11.7)} \qquad \lim_n |S_n h - Sh|_m = 0 \quad \text{for each} \quad h \in C_j.$$

Since C_j is finite dimensional, this implies

$$\lim_n \sup_{|h| \leq 1, h \in spC_j} |Sh - S_n h|_m = 0.$$

See (11E) for criteria on (11.7).

Finally, assume

$$\text{(11.8)} \qquad \text{if } h_n, h \in C_j \quad \text{and} \quad d_F(h_n, h) \to 0, \quad \text{then} \quad h_n \to h \quad \text{in measure.}$$

See Section 3 for criteria on (11.8).

(11.9) PROPOSITION. *Suppose given an increasing sequence of closed convex sets C_j as described in this subsection. Assume* (6.1)–(6.3), (11.5)–(11.8). *Then there exist integers $n_1 < n_2 \cdots$ such that if $D_n = C_j$, $n_j \leq n < n_{j+1}$, then* (6.4)–(6.6) *hold.*

PROOF. Let ${}_jS$, ${}_jS_n$ denote the operators S, S_n restricted to span C_j. Let $\|_j$ denote the operator norm here, so $|{}_jS - {}_jS_n|_j \to 0$ because C_j is finite dimensional. Again because of finite dimensionality of C_j, $|{}_jS_n^{-1}|_j \to |{}_jS^{-1}|_j$. Therefore for each j there exists n_j, $n_j < n_{j+1}$, such that $1/|{}_jS_n^{-1}|_j \geq \frac{1}{2}\, 1/|{}_jS^{-1}|_j$ for all n n_j. Moreover if $h \in \text{sp } C_j$, $|T(n; h)|_L \leq |h| \leq |{}_jS_n^{-1}|_j |{}_jS_n h|_m$. So if D_n is defined to be C_j, $n_j \leq n < n_{j+1}$, then (6.4) holds with $y_n = \frac{1}{2}\, 1/|{}_jS_n^{-1}|_j$, $n_j \leq n < n_{j+1}$.

Next, fix n and $g_0 \in D_n$; suppose $n_j \leq n < n_{j+1}$. There exists $v_j > 0$ such that if $g \in D_n = C_j$ and $|T(n; g) - g_0|_L > r_1^{-1}(y_n)$, then $|FT(n; g) - FT(n; g_0)|_m v_j$. Indeed, suppose not. Then there exists a sequence g_k in C_j such that $|FT(n; g_k) - FT(n; g_0)|_m \to 0$. But then by (11.8) g_k converges to g_0 in measure; since C_j is finite dimensional, g_k converges to g_0 in norm. But then $T(n; g_k)$ converges to $T(n; g_0)$ in norm, and that is a contradiction. Therefore, if z'_j is defined by $v_j = z'_j r_1^{-1}(\frac{1}{2} y_j)$, then (6.5) holds with $z_n = z'_j$ for $n_j \leq n < n_{j+1}$. Finally, it is clear that by taking the n_j even larger, if necessary, the other hypothesis can be satisfied.

(11.10) REMARK. A special case of practical interest occurs when Γ is finite dimensional. Then one naturally can take $C_j = \Gamma$ all j; i.e., under the hypothesis (11.5)–(11.8) one can take $D_n = \Gamma$ in n, and, of course, y_n, z_n are then independent of n.

11D. HYPOTHESES (6.4)–(6.6), $k_n \ll n$. This subsection shows that, under mild conditions, essentially no matter what sequence $D_n \uparrow \Gamma$ is used (even infinite dimensional ones), it is possible to choose k_n so that (6.4), (6.5) hold. To achieve (6.6) it is then a matter of making k_n increase much slower than n. Here are the requisite hypotheses for this development. Assume (6.1)–(6.3). Assume further

(11.11) there is an increasing sequence of subsets $\Theta_n \subset \Theta$, $\Theta_n \uparrow \Theta$ such that $C_n^2 \equiv \int \inf_{\theta \in \Theta_n} F_1^2(\theta; t) m(dt) > 0$,

(11.12) if $g \in D_n$, range $g \subset \Theta_n$.

(11.13) PROPOSITION. *Under these hypotheses,* (6.4), (6.5) *hold; y_n and z_n may be chosen to be constant multiples of C_n/k_n.*

Hypotheses (11.11), (11.12) are easy to satisfy in many examples: one needs only $|F_1(\theta; t)| > 0$ for all θ and a few t. Usually there is enough smoothness so that one may take C_n decreasing as slowly as desired by proper choice of Θ_n. Using this, and making k_n increase very slowly, one can then ensure (6.6); typically there are a great number of ways of arranging this. In the shift model $F(\theta; t) = F(t - \theta)$, where the density f is (say) $\mathcal{N}(0, 1)$, appropriate Θ_n might be $[-b_n, b_n]$ for some increasing sequence B_n; if m places any mass at all near 0 then one could take $C_n = (2\pi)^{-1/2}\exp(-b_n^2/2)$ essentially.

PROOF. The proof is based on the following lemma. Define the compact operator τ mapping L to L by $(\tau g)(s) = \int_0^s g(u)du$. Let $\mathscr{F}_n$ be the sigma field on [0, 1] generated by the intervals $[i/n, (i+1)/n)$ and let $\mathscr{H}_{n,k}$ be the sigma field generated by all step functions $T(n, k; g)$, $g \in L$.

(11.14) LEMMA. (a) *There exists a constant* $C > 0$ *such that if* g *is* $\mathscr{F}_n$ *measurable then* $|\tau g|_L \ge Cn^{-1}|g|_L$. (b) *Let* k_n *be a sequence such that* $k_n/n \to 0$. *There is a constant* $C > 0$ *such that if* g *is* $\mathscr{H}_{n,k_n}$*-measurable then* $|\tau g|_L \ge Ck_n^{-1}|g|_L$. *It is easy to see that the rate* n^{-1} *of* (11.14a) *is best possible.*

To apply this lemma, notice that if t is fixed then

$$A(T(n; k_n; g_0); T(n; k_n; g - g_0))(s, t) = (\tau h_t)(s),$$

where

$$h_t(u) = F_1(T(n; k_n; g_0)(u); t)T(n; k_n; g - g_0)(u).$$

Of course, h_t is $\mathscr{H}_{n,k_n}$-measurable. Therefore, by the lemma,

$$|A(T(n; k_n; g_0); T(n; k_n; g - g_0))(\cdot, t)|_L \ge Ck_n^{-1}\,|h_t|_L$$
$$\ge Ck_n^{-1}\inf_{\theta\in\Theta_n}|F_1(\theta; t)|\,|T(n; k_n; g - g_0)|_L.$$

Integration by $m(dt)$ then completes the verification of (6.4). An entirely similar argument suffices for (6.5).

To understand why the lemma holds, consider the case (a). Let e_i, $1 \le i < n$ be the indicator of the interval $((i-1)/n, i/n]$ and set $g_i = e_{i+1} - e_i$, $1 \le i < n$, $g_n = e_n$. Then the g_ι are linearly independent, span the space of $\mathscr{F} n$ measurable functions, and the lemma holds for each g_ι by direct calculation. Using the Gram-Schmidt orthogonalization, one may show, with some tedium, that the lemma continues to hold for any linear combination of the g_i. The system g_ι was chosen, of course, so that the images τg_i have essentially the same orthogonality properties as g_i. Part (b) may be proved by similar methods.

11.E. HYPOTHESIS (6.7). Roughly speaking, it is clear that (6.7) will hold if $T(n; g)$ is uniformly close to g for $g \in D_n$ and if $A(\cdot, \cdot)$ has sufficient joint continuity. Here is one way to ensure this. By Schwarz, if t is fixed,

$$|A(T(n; g_0); T(n; g))(\cdot, t)|_L$$
$$\le |F_1(T(n; g_0); t)|_L|\,T(n; g) - g|_L + |F_1(T(n; g_0); t) - F_1(g_0; t)|_L|g|_L.$$

Let Θ_n be an increasing sequence of subsets of Θ; assume $g \in D_n$ has its range in Θ_n. Set

$$v_n^2 = \int \sup_{\theta\in\Theta_n}|F_1(\theta; t)|^2 m(dt), \qquad w_n^2 = \int \sup_{\theta\in\Theta_n}|F_2(\theta; t)|^2 m(dt).$$

Then, ignoring a few constants,

$$|A(T(n; g_0); T(n, g)) - A(g_0; g)|_m \le v_n\,|T(n, g) - g|_L + w_n|g\;|_L\,T(n; g_0) - g_0\,|_I.$$

In typical applications, the numbers v_n, w_n can be made to increase as slowly as desired by appropriate choice of Θ_n. The sets D_n can easily be chosen to ensure $\sup_{g\in D_n}|T(n; g) - g|_L \to 0$ as $n \to \infty$. For example, if every $g \in D_n$ has a continuous derivative bounded by M_n, then $|T(n; g) - g|_L \le M_n/n$ and it is then a matter of choosing M_n so that $M_n \uparrow \infty$ and $(v_n + w_n)M_n/n \to 0$, while keeping the norm of elements of D_n from getting big too fast. In the case Γ is finite dimensional, it is particularly simple to guarantee (6.7).

12. Proofs of (3.6), (3.7). We prove here the characterizations (3.6), (3.7) of convergence in the metric d_F of Section 3; proofs of the criteria (3.8), (3.9) are omitted.

PROOF OF (3.6). It is enough to prove that $d_F(g_n, g) \to 0$ implies g_n converges to g in measure, because the reverse implication is trivial. For this, fix s; then the sub probability measures of mass s having distribution function $Fg_n(\cdot; s)$ converge to one with df $Fg(\cdot; s)$. The characteristic functions of these measures therefore converge, leading to

$$\lim \int_0^s \exp\{iag_n(u)\}du = \int_0^s \exp\{iag(u)\}du, \quad \text{all } s, a,$$

since $F(\theta; t) = F(t - \theta)$. The result is then immediate by the following amusing result, whose proof, and the many ponderous generalizations thereof, will be left to the reader.

(12.1) LEMMA. *Let g_n, $n = 1, 2, \cdots$ and g be measurable functions on the Lebesgue unit interval. The following are equivalent*:

(a) *g_n converges to g in measure,*

(b) *g_n converges in distribution to g on every subinterval* $[0, s]$ (i.e., *for each s, if f is any continuous function on* $[0, s]$, *then* $\lim \int_0^s f(g_n(u))du = \int_0^s f(g(u))du$),

(c) *for every continuous f on* $[0, 1]$, *$f(g_n)$ converges in the weak topology of L to $f(g)$.*

PROOF OF (3.7). Suppose $d_F(g_n, g) \to 0$. It will be shown that every subsequence of g_n has a further subsequence that converges in measure to g; since convergence in measure is metric, this will prove the result. Select any subsequence g_{n_1} of g_n. Since $Fg_n(,)$ converges in L_m to Fg, there is then a further subsequence g_{n_2} converging a.e. $dsdm$ and therefore, since Fg is a continuous distribution function, $Fg_{n_2}(s, t) \to F(g(s, t))$ for all s, t. Fix s, so $Fg_{n_2}(s, t) \to F(g(s, t))$ for all t. Since the elements of Γ are L-bounded by hypothesis, the distributions of the g_n are tight. There is, for each s as above, a further subsequence $g_{,3}$, weakly convergent on $[0, s]$ to some g_s. Here weak convergence is in the usual sense of probability distributions; the subsequence n_3 depends on s, a priori. Since $F(\theta; t)$ is continuous in θ

$$\lim_n \int_0^s F(g_{n_3}(u); t)du = \int_0^s F(g_s(u); t)du.$$

Since n_3 is a subsequence of n_2, $\int_0^s F(g_s(u); t)du = \int_0^s F(g(u); t)du$ so by (3.12) $g_s(u) = g(u)$ a.e. on $[0, s]$: i.e., the limit in distribution on $[0, s]$ of any weakly convergent subsequence of g_{n2} is always g. It follows that g_{n_2} converges in distribution to g on each interval $[0, s]$. Appeal to (12.1) completes the proof.

13. Extensions.

(13A) *Multidimensional* Θ. The theory of this paper can be extended to include families $F(\theta; t)$ where $\theta \in \Theta$, a convex subset of R^d. For such Θ, one can define $c_d(n; i; g)$, where $g = (g_1, \cdots, g_d)$, $g_k \in L$, to be the point in R^d given by $(c(n; i; g_1), \cdots, c(n; i; g_d))$. Then the modified regression model consists of independent real random variables $X_{n1}, \cdots, X_{nn}$, where the distribution of X_{ni} is $F(c_d(n; i; g); dx)$. Of course g will have to be chosen so that $c_d(n; i; g) \in \Theta$. Need for this particular extension arises often; it was not treated in the basic development of this paper because it involves substantial notational complication, with no particular gain in insight. For a simple familiar example, consider the "straight line" regression problem where X_{ni} has the form $X_{ni} = a(i/n) + b + Z_i$, with Z_i independent, $N(0, c^2)$, and the numbers a, b, c unknown; the task is to estimate (a, b). This is a special case of a two-dimensional Θ with $F((\theta_1, \theta_2); t) = F((t - \theta_1)/\theta_2)$. The relevant functions $g = (g_1, g_2)$ that arise, of course consist of all g_1 having the form $as + b$ for some real a, b and all g_2 with the form $g_2(s) = c$ for some $c > 0$. This particular model assumes that the variance of each "error" Z_i is the same (but unknown) at every observation. The development of this paper allows as well for a formulation in which the variance c_{ni}^2 of X_{ni} varies arbitrarily with i: evidently it is just a matter of replacing the allowable functions g_2 just mentioned by the set of all elements of L that are positive a.e. Finally, one can even set up the model so that, not only are the variances arbitrary, but the means as well: this means only that the allowable functions g_1 be augmented to all of (or to a dense subset of) L. In each case, the evident analogues of the minimum distance recipes of Section 4 are easily formulated, and, with some personal pain, the asymptotic results can be recovered by the methods of this paper.

(13B) *Dependent observations.* The assumption that the X_{ni} be independent can be weakened. If, under the envisioned dependence, (5.3) still holds, and if independence is a special case of the type of dependence under consideration, then the main results may be recovered. In particular, the proposed estimators will still be robust under small departures from independence.

(13C) *Testing.* The results of this paper automatically suggest tests of various hypotheses. For example, in testing $g \equiv 0$ (i.e., the observations are identically distributed), one naturally would reject the null hypothesis if $\hat{F}_n$ were too far away from $FT(n; g_0)(g_0 \equiv 0)$ in the distance $\|_m$. An asymptotic minimax definition of robustness can be developed for the testing situation, and tests such as the one just proposed may be shown to be robust. Their asymptotic behaviour can be derived from the asymptotic results of this paper. See Millar (1979b), to see one possibility for the location model (among others).

REFERENCES

BERAN, R. J. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5** 445–463.

BERAN, R. J. (1981). Efficient robust estimation in parametric models. *Z. Wahrsch. verw. Gebiete* **57** 73–86.

BICKEL, P. J. (1975). One step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–433.

BICKEL, P. J. (1979). Quelques aspects de la statistique robuste. *École d'Été de St. Flour.* Springer, Berlin. To appear.

BICKEL, P. J. and WICHURA, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.* **42** 1656–1670.

BOLTHAUSEN, E. (1977). Convergence in distribution of minimum distance estimates. *Metrika* **24** 215–217.

CAIROLI, R. and WALSH, J. B. (1975). Stochastic integrals in the plane. *Acta Math.* **4** 111–183.

CHATTERJI, S. D. (1973). Les martingales et leurs applications analytiques. *École d'Été de Probabilités.* Springer Lecture Notes **307** 27–164.

HABERMAN, S. J. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist.* **5** 815–841.

HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berkeley Symp.* **I** 175–194.

HÁJEK, J. and ŠIDÁK, Z. (1967). *Theory of Rank Tests.* Academic, New York.

HUBER, P. (1973). Robust regression. *Ann. Statist.* **1** 799–821.

JOHANSEN, S. (1979). Introduction to the theory of regular exponential families. Univ. Copenhagen Lecture Notes **3**.

KIEFER, J. (1972). Skorokhod embedding of multivariate rv's and the sample d.f. *Z. Wahrsch. verw. Gebiete* **24** 1–35.

KIEFER, J. and WOLFOWITZ, J. (1958). On the deviation of the empiric distribution function of vector chance variables. *Trans. Amer. Math. Soc.* **87** 173–186.

LECAM, L. (1969). Théorie asymptotique de la decision statistique. Les Presses de l'Université de Montreal.

LECAM, L. (1972). Limits of experiments. *Proc. Sixth Berkeley Symp.* **I** 245–261.

MARONNA, R. and YOHAI, V. (1979). Asymptotic behaviour of M estimators for the linear model. *Ann. Statist.* **7** 258–268.

MILLAR, P. W. (1979a). Asymptotic minimax theorems for the sample distribution function. *Z. Wahrsch. verw. Gebiete* **48** 233–252.

MILLAR, P. W. (1979b). Robust tests for statistical hypotheses. Preprint.

MILLAR, P. W. (1981a). Robust estimation via minimum distance methods. *Z. Wahrsch. verw. Gebiete* **55** 73–89.

MILLAR, P. W. (1981b). The minimax principle in asymptotic statistical theory. To appear, *Proc. Ecole d'Ete St. Flour.*

PARR, WILLIAM C. (1980). Minimum distance estimation: a bibliography. *Comm. Statist.*

PARR, W. C. and SCHUCANY, W. R. (1980). Minimum distance and robust estimation. *J. Amer. Statist. Assoc.* **75** 616–624.

POLLARD, D. (1980). The minimum distance method of testing. *Metrika* **27** 43–70.

RIEDER, H. (1981). On local asymptotic minimaxity and admissibility in robust estimation. *Ann. Statist.* **9** 266–277.

SACKS, J. and SPIEGELMAN, C. (1980). Consistent window estimation in non parametric regression. *Ann. Statist.* **8** 240–246.

STONE, C. (1977). Consistent non parametric regression. *Ann. Statist.* **5** 595–645.

WILLIAMSON, M. A. (1980). Weighted empirical type estimation of the regression parameter. *J. Multivariate Anal.* **10**.

WOLFOWITZ, J. (1957). The minimum distance method. *Ann. Math. Statist.* **28** 75–88.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720