TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 286, Number 1, November 1984

A GENERAL APPROACH TO THE OPTIMALITY OF MINIMUM DISTANCE ESTIMATORS

BY
P. W. MILLAR[1]

ABSTRACT. Let Θ be an open subset of a separable Hilbert space, and $\xi_n(\theta)$, $\theta \in \Theta$, a sequence of stochastic processes with values in a (different) Hilbert space B. This paper develops an asymptotic expansion and an asymptotic minimax result for "estimates" $\hat{\theta}_n$ defined by $\inf_\theta |\xi_n(\theta)| = |\xi_n(\hat{\theta}_n)|$, where $|\cdot|$ is the norm of B. The abstract results are applied to study optimality and asymptotic normality of procedures in a number of important practical problems, including simple regression, spectral function estimation, quantile function methods, min-chi-square methods, min-Hellinger methods, minimum distance methods based on M-functionals, and so forth. The results unify several studies in the literature, but most of the LAM results are new. From the point of view of applications, the entire paper is a sustained essay concerning the problem of fitting data with a reasonable, but relatively simple, model that everyone knows cannot be exact.

I. Introduction. This paper presents a rather general approach to the optimality properties of a broad class of estimators of the "minimum distance type". To see what this approach is, let us first consider several important, practical special cases.

Let us begin with the classic case. Suppose $X_1, \ldots, X_n$ are i.i.d. random variables, the distribution of X_i being P_θ, where $\{P_\theta, \theta \in \Theta\}$ is a family of probabilities on the line indexed by Θ, an open subset of R^d. The statistical problem is to "estimate Θ"; as pointed out several times in this paper, the phrase "estimate Θ" as it has been used throughout much of the statistical literature is usually too vague to have a solid statistical meaning. However, for the purposes of these introductory heuristics let us suppose that we know what this means. Let F_θ be the cdf of P_θ; let $\hat{F}_n$ be the empirical cdf. Then $F_\theta, \hat{F}_n$ can be regarded as elements of several varieties of Banach spaces. For example, if μ is a finite measure on the line, $F_\theta, \hat{F}_n$ belong to $L^2(\mu)$ or to $L^1(\mu)$; if F_θ is continuous and a standard smoothing of $\hat{F}_n$ adopted, then $F_\theta, \hat{F}_n$ both belong to a Banach space of real bounded continuous functions with supremum norm. Therefore, let us denote by B the Banach space in which $F_\theta, \hat{F}_n$ are alleged to lie, and by $|\cdot|_B$ the norm of B. The classic minimum distance estimate of θ, based on $\hat{F}_n$, is then the point $\hat{\theta}_n$ that satisfies

$$(1.1) \qquad \inf_\theta |F_\theta - \hat{F}_n|_B = |F_\theta - F_{\hat{\theta}_n}|_B.$$

Received by the editors January 22, 1984.

1980 *Mathematics Subject Classification*. Primary 62F12; Secondary 62F35, 62G20.

Key words and phrases. Local asymptotic minimax, minimum distance estimator, robust estimator, Cramer-von Mises statistic, minimum chi-square, differentiable statistical functional, abstract Wiener space.

[1] Research supported by National Science Foundation Grant MCS80-02648.

Here, and elsewhere in this introduction, we ignore the problems of existence and uniqueness of such a $\hat{\theta}_n$: these difficulties will be treated in subsequent sections.

There are innumerable variants of the recipe (1.1). For example, one could let ϕ_θ be the characteristic function (ch.f.) of P_θ, and $\hat{\phi}_n$ the ch.f. of $\hat{F}_n$, and define another minimum distance estimate $\hat{\theta}_{2n}$ by

$$\inf_\theta |\phi_\theta - \hat{\phi}_n|_B = |\phi_{\hat{\theta}_{2,n}} - \hat{\phi}_n|_B, \tag{1.2}$$

where, as before, there is quite a bit of freedom in the choice of B. As another possibility, one could let F_θ^{-1} denote the quantile function of P_θ: the function on $[0,1]$ defined by $F_\theta^{-1}(t) = \inf\{s\colon F(s) \geq t\}$, and then one could let $\hat{F}_n^{-1}$ be the quantile function of $\hat{F}_n$. Yet another minimum distance estimate, $\hat{\theta}_{3,n}$, of θ is then possible via the recipe

$$\inf_\theta |\hat{F}_n^{-1} - F_\theta^{-1}|_B = |\hat{F}_n^{-1} - F_{\hat{\theta}_{3,n}}^{-1}|_B \tag{1.3}$$

Again there is much freedom in the choice of B. All of these estimators have been considered in the literature; for some new ones in a similar vein, which have additional interesting properties, see §13. More complexity is possible—and even useful: one can replace the norm $|\ |_B$ by a family of norms $|\ |_\theta$; this device leads to the well-known weighted Cramer-von Mises estimates.

The illustrations of the preceding paragraph dealt with estimation in a certain i.i.d. framework. There are other situations, of great practical interest, which share exactly the same abstract structure. For a first example, let $X_1, \ldots, X_n$ be a stationary Gaussian sequence with mean 0, spectral measure P_θ, where $\theta \in \Theta$, an open subset of R^d. Again one wants to "estimate θ". Let

$$\hat{F}_n(t) = (2\pi n)^{-1} \int_0^{+} \left| \sum_1^n X_0 e^{-iyk} \right|^2 dy.$$

Let $F_\theta(t)$ be the cumulative of the spectral measure P_θ and bring in an appropriate Banach space B as in the preceding example. Then a minimum distance estimate $\hat{\theta}_n$ of θ is defined by

$$\inf_\theta |F_\theta - \hat{F}_n|_B = |F_{\hat{\theta}_n} - \hat{F}_n|_B. \tag{1.4}$$

As one further example, consider a "simple regression" model. Let $X_{n1}, \ldots, X_{nn}$ be independent random variables. Assume a distribution F is given and there are *known* numbers c_{ni}, $1 \leq i \leq n$, such that $\{X_{ni} - \theta c_{ni},\ 1 \leq i \leq n\}$ is an i.i.d. (F) sequence for some θ. Again the problem is to estimate θ. Define a random function

$$\xi_n(\theta; t) = \sum d_{ni}[I\{X_{ni} \leq t + \theta c_{ni}\} - F(t)],$$

where $\{d_{ni},\ 1 \leq i \leq n\}$ is a sequence of real numbers. Bring in, yet again, some Banach space B, so that $\xi_n(\theta;\cdot) \in B$ for each θ. A minimum distance estimate of θ, $\hat{\theta}_n$, is then defined by

$$\inf_n |\xi_n(\theta;\cdot)|_B = |\xi_n(\hat{\theta}_n;\cdot)|_B. \tag{1.5}$$

The estimator (1.5) has been partially analyzed in the literature. The estimate (1.4) appears to be new (see §§10, 11).

One of the main points of this paper is that, under mild, natural conditions, all of the estimates described above—and many more—share a common, very simple abstract structure asymptotically. This common structure enables one to deduce immediately the asymptotic normality and, just as important, a strong asymptotic optimality.

The abstract structure is of the following sort. One is given a sequence $\xi_n(\theta)$ of B-valued stochastic processes ('time' parameter is $\theta \in \Theta$). The statistical problem is to compute $\hat{\theta}_n$ satisfying

$$\inf_\theta |\xi_n(\theta)|_B = |\xi_n(\hat{\theta})|_B.$$

In (1.1), $\xi_n(\theta)$ could be $F_\theta - \hat{F}_n$; in (1.2), $\xi_n(\theta)$ could be $\hat{\phi}_n - \phi_\theta$ (or $\sqrt{n}(\hat{\phi}_n - \phi_\theta)$), and so forth. Under reasonable smoothness conditions on the sample functions $\theta \to \xi_n(\theta)$, it is possible to give a very simple characterization of $\xi(\hat{\theta}_n)$, a characterization that leads swiftly in all of the applications to an asymptotic normality result for $\hat{\theta}_n$. These matters are discussed in §§2, 3. There is a long history of minimum distance estimation; see Parr for a survey. The development here draws on the analyses of Bolthausen, Millar, Pollard and Wolfowitz.

The optimality properties of such minimum distance estimators are harder to describe. To get a feel for it, return to the classic case (1.1). Let us suppose $X_1, \ldots, X_n$ are i.i.d. G, but G is *presumed* to be possibly different from any of the P_θ's. This fact notwithstanding, the basic statistical goal is defined to being that of fitting the "model" $\{P_\theta\}$ as "best as one can" to the data. There are several different situations in which one might desire to do this. One possibility is to assume that the $\{P_\theta\}$ model would have been fairly accurate, except somewhere along the line the data was subjected to "contamination" (thus altering the distributions slightly); such contamination could be due to roundoff errors, clerical errors, and so forth. Alternatively, one might be certain that model is inaccurate, but not too bad, and desire to use it because (say) it has an appealing simplicity that aids in getting a feel for the basic shape of the data. The precise notion of 'best fit' to be employed should be determined by your statistical goals. If you want to estimate a few probabilities, then (1.1) might be appropriate; if you wanted to estimate a few quantiles, perhaps (1.3) would be better. There are, evidently, a large number of choices, and a fair portion of this paper is devoted to explaining some of the consequences of a particular choice. The optimality results of this paper are elaborated in a framework where it is agreed that the statistical goal is one of "best fit"; within such a framework, estimators of the type suggested by $\hat{\theta}_{1n}, \hat{\theta}_{2n}, \hat{\theta}_{3n}$, and many others, are shown to have a desirable stability property, called local asymptotic minimaxity (LAM). Roughly speaking, the optimality property asserts that estimators do not deteriorate when the actual data distribution departs somewhat from those given by the model: i.e., the minimum distance estimators are robust.

For the classic case one can formulate the estimation problem to be solved as follows: similar formulations extend to all other applications. The observations $X_1, \ldots, X_n$ are assumed i.i.d., with unknown distribution G. Define a functional $G \to \theta(G)$ by

$$\inf_\theta |G - F_\theta|_B = |G - F_{\theta(G)}|_B. \tag{1.6}$$

Then the goal is to estimate $\theta(G)$, where G is the actual data distribution. Thus, we estimate the $\theta \in \Theta$ that ensures the "best fit" of the model to the data, where (in the present case) 'best fit' is defined by (1.6). We emphasize again, that there are many choices for the notion of 'best fit', and the one you actually decide on depends upon what use you are trying to make of the data. However, given that your goal is that of best fit, it is quite irrelevant to evaluate your estimates on the basis of the criterion of 'least asymptotic variance': estimators with small asymptotic variance will *not* necessarily ensure your announced goal of 'best fit'. This should be kept in mind while perusing the optimality results below.

The paper is organized as follows. §§2, 3 give the basic asymptotic expansion that ensures (in all the examples treated) asymptotic normality of the minimum distance estimators. §§4, 5 give, in abstract form, the basic optimality result. §§6–14 discuss a wide variety of applications. You may complain initially about the relatively high level of abstraction in the basic development—especially §§4, 5. On the other hand, once you read through the large and varied selection of useful applications, I think you will find it not unreasonable. §15 gives a couple of proofs omitted from §§2, 3.

The development of this paper unifies several results in the literature and provides a fair number of new results. The LAM results in the sections concerning spectral functions, quantile functions, min-Hellinger methods, the variety of methods in §13, the simple regression problem, possibly also §8, appear here for the first time. In most of these applications (except §§11, 13), asymptotic normality was either known or part of the folklore.

II. A basic asymptotic expansion. This section contains the basic asymptotic representation of minimum distance estimators which will be specialized in the applications of subsequent sections. That this level of abstraction is necessary can be seen quickly on reviewing at the same time (e.g.) the spectral function example and the regression example in §1.

To describe the result, let Θ be a separable Hilbert space and fix $\theta_0 \in \Theta$. Let $(\Omega_n, \mathcal{F}_n, P_n)$ be a sequence of probability spaces, and let $\{\xi_n(\theta, \omega),\ \theta \in \Theta\}$ be a sequence of stochastic processes on $(\Omega_n, \mathcal{F}_n)$; here $\omega \in \Omega_n$. The processes $\xi_n(\theta, \omega)$ are assumed to have their values in a *separable Hilbert space B with norm* $|\ |_B$. It is important that, so far, no hypotheses are placed on the probabilities P_n; in some useful applications (which show the Fréchet differentiability of certain functionals), P_n is, for each n, concentrated at a single point.

Fix $\theta_0 \in \Theta$. Introduce the following hypotheses and definitions.

(2.1) *Hypothesis of identifiability.* As $c \uparrow \infty$, then, under P_n,

$$\inf_n \inf_{|\theta - \theta_0| > c} |\xi_n(\theta; \cdot)|_B \to \infty.$$

Measurability problems here can be avoided by taking a separable version of the process $\xi_n(\theta, \cdot)$ (cf. Doob (1953)). Many refinements of this hypothesis are possible—with substantial increase in the level of tedium; for example, the inf need only be taken over "sufficiently large" n, an easy reformulation left to the reader.

The second hypothesis is

(2.2) *Differentiability.* There exists a continuous linear map $T\colon \Theta \to B$ and a sequence of real numbers δ_n such that, for all c,

$$\sup_{|\theta-\theta_0|<c} |\xi_n(\theta) - \xi_n(\theta_0) - \delta_n T(\theta - \theta_0)|_B = o_{P_n}(1)$$

under P_n. Here ξ_n is random (i.e., depends on $\omega \in \Omega$) but T is assumed nonrandom.

When $\Theta = R^d$, one can endow Θ with the usual basis $\{\varepsilon_1, \dots, \varepsilon_d\}$, so $\theta = \sum \theta_i \varepsilon_i$. Set $\eta_i = T(\varepsilon_i)$, $\eta = (\eta_1, \eta_2, \dots, \eta_d)$. Then

$$T(\theta) = \langle \theta, \eta \rangle = \sum \theta_i \eta_i. \tag{2.2a}$$

(2.3) DEFINITION. The derivative T is nonsingular if, for all θ, $|T(\theta)| \geq b|\theta|$ for some $b > 0$.

In the case that $\Theta = R^d$, this assumption is equivalent to the condition that the η_i in (2.2a) be linearly independent. The third assumption is

(2.4) *Boundedness.* Under $\{P_n\}$ the B-valued random variables $\xi_n(\theta_0)$ are norm bounded in probability: for each $\varepsilon > 0$ there exists c such that

$$P_n\{|\xi_n(\theta_0)|_B > c\} \leq \varepsilon$$

for all n. Alternatively, one could simply say that the P_n-distributions of the real random variables $|\xi_n(\theta_0)|_B$ are *tight.*

In many examples a stronger hypothesis holds.

(2.5) *Convergence.* Under P_n, the B-valued random elements $\xi_n(\theta_0)$ converge weakly on B to a random element W.

This hypothesis, of course, implies (2.4); but in several applications it is *not* satisfied.

(2.6) DEFINITION. Define B_η to be the subspace of B spanned by $\{T(\theta), \theta \in \Theta\}$ and define π to be *orthogonal projection* of B to B_η.

Define the *minimum norm* estimate $\hat{\theta}_n$ to be *any* point $\hat{\theta} \in \Theta$ satisfying

$$\inf |\xi_n(\theta)|_B = |\xi_n(\hat{\theta}_n)|_B. \tag{2.7}$$

There are, at this point, problems of existence of $\hat{\theta}_n$; under assumptions (2.1), (2.2), (2.4) these problems disappear asymptotically. The worried reader can instead consider *any* $\hat{\theta}_n$ that achieves the inf in (2.6), with an error of n^{-2}: $\hat{\theta}_n$ should satisfy $\inf_\theta |\xi_n(\theta)|_B \geq |\xi_n(\hat{\theta}_n)|_B - n^{-2}$. Now there are no existence problems; however, since asymptotically we *do* have existence and uniqueness, we shall, for notational simplicity, proceed as though the inf were achieved.

The basic asymptotic expansion can now be stated.

(2.8) THEOREM. *Assume the hypotheses of identifiability, differentiability, and boundedness. Define $\hat{\theta}_n$ by* (2.7) *and π by* (2.6). *Then*

$$\xi_n(\hat{\theta}_n) = (1 - \pi)\xi_n(\theta_0) + o_{P_n}(1).$$

As one might expect, a result this general has a crude proof based only on elementary considerations. This fact does not vitiate its utility. To see heuristically that the form of the result is reasonable, note that (2.1), (2.4) suggest that $\hat{\theta}_n$, which minimizes the norm, must be in a vicinity of θ_0. This being so, the differentiability assumption (2.2) implies

$$\begin{aligned}|\xi_n(\theta)|^2 &= |\xi_n(\theta) - \xi_n(\theta_0) + \xi_n(\theta_0)|^2 \doteq |\xi_n(\theta_0) + \delta_n T(\theta - \theta_0)|^2 \\ &= |\pi\xi_n(\theta_0) + \delta_n T(\theta - \theta_0)|^2 + |(1-\pi)\xi_n(\theta_0)|^2\end{aligned}$$

so that the minimum in θ should be at $\hat{\theta}_n$ satisfying $\pi\xi_n(\theta_0) + \delta_n T(\hat{\theta}_n - \theta_0) = 0$. This means that

$$\text{(2.9)} \qquad \begin{aligned}\xi(\hat{\theta}_n) &= \xi_n(\theta_0) + \xi_n(\hat{\theta}_n) - \xi_n(\theta_0) \\ &\doteq \xi_n(\theta_0) + \delta_n T(\hat{\theta}_n - \theta_0) = \xi_n(\theta_0) - \pi\xi_n(\theta_0),\end{aligned}$$

from which the result follows.

This result may be translated into a result involving $\hat{\theta}_n$ itself. To develop such a result, assume T is *nonsingular.*

(2.10) COROLLARY. *Assume, in addition to the hypotheses of* (2.8), *that T is nonsingular. Then*

$$\text{(2.11)} \qquad \delta_n(\hat{\theta}_n - \theta_0) = -T^{-1} \circ \pi\xi_n(\theta_0) + o(1).$$

The corollary is immediate from (2.8).

The minimum point $\hat{\theta}_n$ is *asymptotically* unique in the sense that if $\tilde{\theta}_n$ is any other such point (at time n) then $|\hat{\theta}_n - \tilde{\theta}_n| = o(1)$.

The proof of (2.8) appears in §15. To facilitate subsequent applications, let us record the following refinements.

(2.12) *Evaluation of* $T^{-1} \circ \pi$. Assume T is nonsingular. If $x \in B$, x is the unique element of B of the form $\pi x = T\theta_x$ ($\theta_x \in \Theta$) satisfying

$$\inf_\theta |x - T\theta|_B^2 = |x - \pi x|_B^2 = |x - T\theta_x|_B^2.$$

Using the fact that at the minimum point the functional $\theta \to |x - T\theta|^2$ must have derivative 0, together with the form of the derivative at θ as $2\langle T^*x - T^*T\theta, \cdot\rangle$, one finds

$$\text{(2.13)} \qquad \theta_x = K^{-1}T^*x, \quad K = T^*T, \qquad \pi x = TK^{-1}T^*x,$$

where T^* is the adjoint of T. The operator K^{-1} exists and is continuous, because of the nonsingularity hypothesis.

If $\xi_n(\theta_0)$ converges in B to a Gaussian random element W with mean 0 and covariance operator S, then evidently, by (2.11),

$$\text{(2.14)} \qquad \delta_n(\hat{\theta}_n - \theta_0) \Rightarrow K^{-1}T^*W,$$

a Gaussian random element with mean 0 and covariance operator

$$\text{(2.15)} \qquad \Gamma S \Gamma^*, \qquad \Gamma = K^{-1}T^*.$$

In all applications but one, Θ is finite dimensional. To get a more explicit description in this case of (2.14), (2.15), adopt the representation of T given by (2.2a):

$$\text{(2.16)} \qquad T\theta = \sum \theta_i\eta_i.$$

Also, in all applications, B is an $L^2(\mu)$ space of R^d. Assuming this framework, suppose $\xi_n(\theta_n)$ converges in L^2 to a Gaussian process W, where $EW(s) = 0$, $EW(s)W(t) = R(s,t)$, s,t in R^d. Define a matrix $A = (A_{ij})$ by

$$(2.17) \qquad A_{ij} = E\langle \eta_i, W\rangle_B \langle \eta_j, W\rangle_B = \iint \eta_i(s)R(s,t)\eta_j(t)\,\mu(ds)\,\mu(dt),$$

and a matrix $C = (C_{ij})$ by

$$(2.18) \qquad C_{ij} = \langle \eta_i, \eta_j\rangle_B.$$

Then C^{-1} exists by nonsingularity. Specialization of (2.14), (2.15) to the present finite-dimensional case leads easily to

$$(2.19) \qquad \delta_n(\hat{\theta}_n - \theta_0) \Rightarrow Z,$$

where

$$(2.20) \qquad Z \text{ is } N(0, C^{-1}AC),$$

III. Expansion for processes scaled by $n^{1/2}$. A number of the examples of this paper involve a particular case of the result of §2: processes scaled by $n^{1/2}$. For ease of subsequent applications, we single out this case here.

Here Θ is an open subset of a Hilbert space, $\{P_n\}$ is a sequence of probabilities on $(\Omega_n, \mathcal{F}_n)$, and B is a separable Hilbert space. For each θ, let $\xi_n(\theta,\omega)$ be a B-valued random variable. Define $\hat{\theta}_n$ by

$$(3.1) \qquad \inf_{\theta\in\Theta} |\xi_n(\theta)| = |\xi_n(\hat{\theta}_n)|.$$

Fix $\theta_0 \in \Theta$ and introduce the following variants of the hypotheses of §2.

(3.2) *Identifiability.* For every $\varepsilon, c > 0$ there exists $\delta > 0$ such that

$$P_n\left\{\omega : \inf_{\theta:|\theta-\theta_0|>c} |\xi_n(\theta,\omega) - \xi_n(\theta_0,\omega)| > \delta\right\} \geq 1 - \varepsilon$$

for all n.

In several important applications, $\xi_n(\theta) - \xi_n(\theta_0)$ is free of ω and n, so the formulation is then much simpler.

(3.3) *Boundedness.* Under the probabilities P_n the B-valued random variables $n^{1/2}\xi_n(\theta_0)$ are norm bounded in probability. As in §2, there is a stronger hypothesis, satisfied in most examples:

(3.4) *Convergence.* The B-valued random variables $n^{1/2}\xi_n(\theta_0)$ converge in distribution to a B-valued random variable W.

The third hypothesis is

(3.5) *Differentiability.* There exists a continuous linear operator $T = T_{\theta_0}$ from span Θ to B such that

$$\xi_n(\theta,\omega) = \xi_n(\theta_0,\omega) + T(\theta - \theta_0) + o_P(|\theta - \theta_0|).$$

The definition that T be nonsingular is the same as in §2.

As in §2 let B_η be the subspace of B spanned by $T(\Theta)$ and let π be the orthogonal projection from B to B_η.

(3.6) THEOREM. *Assume identifiability, boundedness, and differentiability* ((3.2), (3.3), (3.5)) *with nonsingular derivative. Define* $\hat{\theta}_n$ *to be the minimum norm estimate* (3.1). *Then with probability approaching* 1 *as* $n \to \infty$, $\hat{\theta}_n$ *exists and is unique. Moreover,*

$$\xi_n(\hat{\theta}_n) = (1-\pi)\xi_n(\theta_0) + o(n^{-1/2}),$$
$$\hat{\theta}_n - \theta_0 = -T^{-1} \circ \pi \circ \xi_n(\theta_0) + o(n^{-1/2}).$$

If, in addition, the hypothesis of convergence holds,

$$n^{1/2}[\xi_n(\hat{\theta}_n) - \xi_n(\theta_0)] \Rightarrow \pi \circ W \quad \text{in } B,$$

$$n^{1/2}[\hat{\theta}_n - \theta_0] \Rightarrow -T^{-1} \circ \pi \circ W \quad \text{in } R^d.$$

As in §2, if W is Gaussian, so are the limits just mentioned and the evaluation of $T^{-1} \circ \pi$ can be found in (2.12).

Since we are interested in this paper *mainly* in the structural aspects of the problems treated, we use the more explicit evaluation (2.12) only in a few cases.

To prove the result just stated, let $\xi_n(\theta)$ be given and define

$$\xi_n^0(\theta) = n^{1/2}[\xi_n(\theta_0 + \theta n^{-1/2})].$$

With modest effort, one sees that the hypotheses of (2.8), (2.11) hold for ξ_n^0; then application of the result of (2.8) yields (3.6).

IV. The basic asymptotic minimax theorem. This section recalls the basic Hajek-Le Cam asymptotic minimax theorem, with the purpose of applying it efficiently to the optimality considerations of subsequent sections. This result is given in (4.9), after preliminaries on the theory of convergent experiments and on abstract Wiener spaces.

Let Θ be an arbitrary index set, and let $\{P_\theta,\ \theta \in \Theta\}$ be a family of probabilities on a measure space $(S, \mathcal{S})$. The triple $E = \{P_\theta, (S, \mathcal{S}),\ \theta \in \Theta\} = \{P_\theta\}$ is a *statistical experiment.*

Suppose, temporarily, that Θ is *finite*, $\Theta = \{\theta_1, \dots, \theta_d\}$. The *canonical measure* μ_E of such an experiment E is the measure on the unit simplex of R^d given by the following recipe: if $M = \sum P_\theta$ and $m = (m_1, \dots, m_d)$, where $m_i = dP_{\theta_i}/dM$, then μ_E is the distribution of the vector m. Suppose $E^n = \{P_\theta^n,\ \theta \in \Theta\}$ is a sequence of experiments indexed by Θ, assumed *finite*. Then E^n *converges* to E if μ_{E^n}, the canonical measure of E^n, converges weakly in R^d to μ_E. Statistical significance of canonical measures was pointed out by Blackwell (1953); the notion of convergence of experiments just defined is due to Le Cam (1964).

Next, suppose Θ is completely arbitrary. Let E^n, E be statistical experiments indexed by Θ. E^n *converges* to E if, for *every finite* subset $\Theta_0 \subset \Theta$, the statistical experiments $\{P_\theta^n,\ \theta \in \Theta_0\}$ converge to $\{P_\theta,\ \theta \in \Theta_0\}$ in the sense just defined for *finite* Θ_0. Le Cam (1964) has given several beautiful equivalences for this mode of convergence.

It is extremely painful to check convergence of canonical measures in general. In this respect, the following lemma is of some help; indeed, all of the convergence results used below can be checked by means of this lemma.

(4.1) LEMMA. *Let E^n, E be statistical experiments indexed by Θ. Suppose there exists $\theta_0 \in \Theta$ such that each P_θ is absolutely continuous with respect to P_{θ_0} and P_θ^n is contiguous to $P_{\theta_0}^n$. For a finite subset Θ_0 of Θ, let μ_{00} denote the distribution of the vector $\{dP_\theta/dP_{\theta_0} : \theta \in \Theta_0\}$ under P_{θ_0}, and define μ_{00}^n similarly for E^n. If, for each Θ_0, μ_{00}^n converges weakly to μ_{00}, then E^n converges to E.*

(4.2) REMARK. Many applications proceed by showing that the distribution of $\log dP_\theta^n/dP_{\theta_0}^n$ converges to that of $\log dP_\theta/dP_{\theta_0}$. Thus, in many statistical applications, the convergence theory of Le Cam is equivalent to certain simple convergences of the likelihood ratios.

The following example appears in some of the applications and is an important example of convergence experiments; it can be checked by routine use of (4.1).

(4.3) EXAMPLE. Let F be a fixed distribution on the line with density f with respect to Lebesgue measures. Let h_0 be a fixed real function and define, for real θ,

$$f(\theta h_0; x) = f(x)[1 + \theta h_0(x)].$$

Assume (a) $\int f(x)h_0(x)\,dx = 0$, and (b) $f(\theta h_0; x) \geq 0$ for all small θ. Then, for all small θ, $f(\theta h_0; \cdot)$ is a probability density. Let P_θ^n be the n-fold product measure of $f(\theta n^{-1/2} h_0; \cdot)$; this will be defined for all large n. Then $\{P_\theta^n,\ \theta \in R'\}$ *converges* to $\{P_\theta,\ \theta \in R^1\}$, where P_θ is the measure on the line with mean $\theta|h_0|$ and variance 1, where $|h_0|^2 = \int h_0^2(x)f(x)\,dx$.

For our applications, a modest extension of Example (4.3) is required. To describe it let H be a separable Hilbert space, B a separable Banach space, and τ a mapping of H to B that is continuous,linear, and one-to-one. Let Q_0 be the cylinder measure on B with characteristic functional

$$\phi(m) = \exp\{-|\tau^* m|_H^2\}, \tag{4.4}$$

where $m \in B^*$, the dual of B, τ^* is the adjoint of τ, and $|\ |_H$ is the norm of H. *Assume* that Q_0 is countably additive on the Borel sets of B. That is, the triple (τ, H, B) is an *abstract Wiener space*. Slightly more detail giving the statistical relevance of such a construct may be found in Millar (1979). Here are two examples.

(4.5) EXAMPLE. $H = R^k$, $\tau h = \Gamma h$, where Γ is a nonsingular matrix, and $B = R^k$. Then Q_0 is the normal distribution on R^k with mean 0 and covariance $\Gamma\Gamma^t$.

(4.6) EXAMPLE. Let F be a probability on the line with density f with respect to m. Define

$$H = \left\{ h\colon \int h^2 f\,dm < \infty,\ \int hf = 0,\ \text{support } h \subset \text{support } f \right\};$$

then H is a Hilbert space with inner product $\langle h_1, h_2 \rangle = \int h_1 h_2 f\,dm$. Define τ, a mapping of h to real functions on the line, by

$$(\tau h)(t) = \int^t h(s)f(s)\,m(ds).$$

Let μ be a finite measure on the line; then $\tau h \in L^2(\mu)$. Define B to be the closure in $L^2(\mu)$ of τH. Let Q_0 be the distribution of the stochastic process $\{W^0(F(t))\}$, where $\{W^0(s),\ 0 < s < 1\}$ is the usual Brownian bridge on $[0, 1]$. Then (τ, H, B)

is an abstract Wiener space since Q_0 is countably additive on B. This is the key example for most of the applications.

(4.7) DEFINITION. Let (τ, H, B) be an abstract Wiener space and Q_0 the measure on B with ch.f (4.4). Define measures Q_h, $h \in H$, on B by $Q_h(A) = Q_0(A - \tau h)$. The statistical experiment $\{Q_h,\ h \in H\}$ is called the standard Gaussian shift family for (τ, H, B).

For checking convergence of experiments to some standard Gaussian shift experiment $\{Q_h\}$ by the method of Lemma (4.1), it is necessary to know that, if $h_1, \ldots, h_d \in H$, then the vector $\{\log dQ_{h_i}/dQ_0 \colon 1 \leq i \leq d\}$ has, under Q_0, a normal distribution with mean zero and a covariance matrix whose $(i-j)$th entry is $\langle h_i, h_j \rangle_H$, where $\langle\ ,\ \rangle_H$ is the inner product of H. See Millar (1979) for a bit more detail on this point.

Here is an important example of statistical experiments E^n converging to a Gaussian shift family.

(4.8) EXAMPLE. Fix $F, f, (\tau, H, B)$ of Example (4.6). For $h \in H$, define $f(h; x) = f(x)[1 + h(x)]$. Denote by $F(h; t)$ the cdf of $f(h; x)$; note that $Fh = F + \tau h$. Let P_h^n be the n-fold product measure of $f(n^{-1/2}h; x)$. Let H_0 be the (dense) subset of $h \in H$ such that $f(n^{-1/2}h)$ is a probability for all sufficiently large n. Then $\{P_h^n \colon h \in H_0\}$ converges to $\{Q_h,\ h \in H_0\}$, the Gaussian shift experiment of (τ, H, B). This may be checked by the method of Lemma (4.1), for example.

We are now ready to describe the required special case of the asymptotic minimax theorem. Let (τ, H, B) be an abstract Wiener space, with $\{Q_h\}$ its standard Gaussian shift experiment. Let H_0 be a dense subset of H, and suppose $E^n = \{P_h^n,\ h \in H_0\}$ is a sequence of experiments converging to $E = \{Q_h,\ h \in H_0\}$. We shall take B as decision space—i.e., all statistical procedures to be considered below shall be B-valued. A nonnegative function l on B is *subconvex* if it is lower semicontinuous, symmetric ($l(x) = l(-x)$) and $\{x \colon l(x) \leq c\}$ is convex for every $c > 0$. For example, if $|\ |_B$ is the norm of B and g is an increasing function, $l(x) = g(|x|_B)$ is subconvex. A *loss function* on $B \times H_0$ shall be defined by $l(x - \tau h)$ if $x \in B$, $h \in H$.

(4.9) THEOREM. *With the assumptions of the preceding paragraph,*

$$\liminf_{n} \inf_{T_n} \sup_{h \in H_0} \int l(T_n - \tau h)\, dP_h^n \geq \int l(x)\, P_0(dx),$$

where the inf is taken over all B-valued estimators available at time n.

Applications to the local asymptotic minimax theory require a slight variant of this result. For each c let N_c be a convex subset of H_0, $N_c \subset N_{c'}$ if $c < c'$, and $\bigcup N_0 = H_0$. Then, with the assumptions of Theorem (4.9),

$$\lim_{c \uparrow \infty} \liminf_{n} \inf_{T_n} \sup_{h \in N_c} \int l(T_n - \tau h)\, dP_h^n \geq \int l(x)\, P_0(dx). \tag{4.10}$$

More abstract and better versions of these results, together with proofs, may be found in Millar (1979) and, of course, Le Cam (1972).

V. A general description of the LAM structure. Estimators of the type introduced in §2 typically have a strong optimality property. This section gives a general description that applies to all of the applications in this paper.

A description of this property involves several ingredients. First, suppose:

(5.1) (τ, H, B) is an abstract Wiener space with $\{Q_h\}$ its Gaussian shift family.

(5.2) For a dense set $H_0 \subset H$, there are statistical experiments $\{P_h^n \colon h \in H_0\}$ converging to $\{Q_h,\ h \in H_0\}$.

For each P_h^n, let

(5.3) $\{\xi_n(\theta, P_h^n) \colon \theta \in \Theta\}$ be a stochastic process with values in B_2, a Hilbert space.

In many applications, $B = B_2$. Let

(5.4) $\theta_{n,h}$ be the point in Θ that achieves

$$\inf_\theta |\xi_n(\theta, P_h^n)|_{B_2} = |\xi_n(\theta_{nh}, P_h^n)|_{B_2}.$$

Of course, θ_{nh} typically will be random. The statistical problem will be to estimate θ_{nh} when the measure P_h^n, $h \in H_0$, is "unknown". More precisely, let g be an increasing function on $[0, \infty)$; define the *loss* when the guess $\tilde{\theta}_n$ is made and when P_h^n is 'true' by

$$g(|\tilde{\theta}_n - \theta_{nh}|), \tag{5.5}$$

where $|\cdot|$ is the norm of R^d.

The *risk* of an estimate $\tilde{\theta}_n$ is then

$$\int g(|\tilde{\theta}_n - \theta_{nh}|)\, dP_h^n. \tag{5.6}$$

Fix now a point $\theta_0 \in \Theta$. *Assume* that the processes $\xi_n(\theta, P_h^n)$ satisfy

$$\xi_n(\theta_0, P_h^n) = \Psi_n + V \circ \tau h + o(1), \quad \forall h \in H_0, \tag{5.7}$$

where

$$V \colon B \to B_2 \tag{5.8}$$

is bounded linear, and Ψ_n is a B_2-valued random variable *independent of h*. In many applications, $B = B_2$ and V is the identity.

Assume further that, for each h, $\xi_n(\theta, P_h^n)$ satisfies the identifiability, differentiability, and boundedness hypotheses of §2 (under the sequence of measures $\{P_h^n\}$), with derivative η independent of h. Let

$$\pi = \text{projection in } B_2 \text{ to } B_\eta \tag{5.9}$$

where B_η is the subspace spanned by the components of η (see (2.6)). Here are two examples which illustrate the foregoing set-up—in particular, (5.7).

(5.10) EXAMPLE (CLASSIC CASE). Let $\{P_\theta,\ \theta \in \Theta\}$ be a parametric family of distributions on the line; assume each P_θ has density $f(\theta; x)$ w.r.t. some sigma-finite measure m. Fix $\theta_0 \in \Theta$ and define densities $f(h; x)$ by $f(h; x) = f(\theta_0; x)[1 + h(x)]$, where $h \in H_0$ is given in (4.6), (4.8) (with $F = P_{\theta_0}$ there). Define P_h^n as in (4.8) (it is the product of $f(hn^{-1/2}))$ and (τ, H, B) as in (4.6). If $F(h; t), F(\theta; t)$ denote cdf's of $f(h; \cdot), f(\theta; \cdot)$, define

$$\xi_n(\theta, P_h^n) = n^{1/2}[F(n^{-1/2}h) - F(\theta_0 + n^{-1/2}\theta)]$$

(the minor abuse of notation here and elsewhere will cause no confusion.) Since

$$F(n^{-1/2}h) = F(\theta_0) + n^{-1/2}\tau h,$$

$\xi_n(\theta, P_h^n) = \xi_n(\theta, P_0^n) + \tau h$ for *every* θ, so (5.8) holds with V the identity. This example is a special case of the situation in §6, where V, in general, is *not* the identity; the fact that the hypotheses of differentiability, etc., hold for $\xi_n(\theta, P_h^n)$, as defined here, is treated in greater generality in §6. But this example furnishes a canonical example of a $\xi_n(\theta, P_h^n)$-process (the measures P_h^n are relevant for checking out the basic hypotheses of differentiability, boundedness, etc., as required for (5.12)).

If $\tilde{\theta}_n$ is an estimate of θ here, let $\tilde{\theta}_{1n} = \theta_0 + n^{-1/2}\tilde{\theta}_n$; define θ_{1nh} similarly. Typically one wants to estimate θ_{1nh} instead; in this case the loss function has the more familiar form

$$g(n^{1/2}|\tilde{\theta}_{1n} - \theta_{1nh}|) = g(|\tilde{\theta}_n - \theta_{nh}|).$$

See the end of this section for a general formulation under "$n^{1/2}$-scaling".

(5.11) EXAMPLE (SIMPLE REGRESSION). (Refer to §1 for a brief description of the relevant model.) Fix a distribution F on the line and define (τ, H, B) as in (4.6). Let $\{d_{ni}\}$ satisfy $\sum_1^n d_{ni}^2 = 1$. Let P_h^n be the product measure $\prod_1^n f(d_{ni}h; x_i)$, where $f(h; \cdot)$ is defined in (4.8). For real θ let

$$\xi_n(\theta, P_h^n)(t) = \sum d_{ni}[I\{X_{ni} \leq t + c_{ni}\theta\} - F(d_{ni}h; t)],$$

where $F(h; \cdot)$ is the cdf of $f(h; \cdot)$. Since $F(d_{ni}h) = F(o) + d_{ni}\tau h$ and $\sum d_{ni}^2 = 1$, ξ_n here satisfies (5.8). Again, the form of P_h^n is used to check the differentiability, et al.

We are now ready for the basic LAM lower bound. Define π by (2.6). Let N_c be any convex subsets of H_0, $N_c \uparrow H_0$. The sets N_c are called *neighborhoods.*

(5.12) THEOREM (LAM LOWER BOUND).

$$\lim_{c\uparrow\infty} \liminf_n \inf_{\tilde{\theta}_n} \sup_{h\in N_c} \int g(\delta_n|\tilde{\theta}_n - \theta_{nh}|)\, dP_h^n \geq \int g(|T^{-1} \circ \pi \circ Vx|\, Q_0(dx) = Eg(Z).$$

Here the infinium is computed over all estimates $\tilde{\theta}_n$ of θ_{nh} and

$$(5.13) \qquad Z = T^{-1} \circ \pi \circ V \circ X,$$

where X is a B-valued random variable with distribution Q_0; Q_0 was given in (5.1). Of course Z has a normal distribution on R^d with mean 0; its covariance can be calculated using (2.20).

(5.12) gives an asymptotic lower bound to the minimax risk (when it is defined by (5.6)). The chief difficulty in applying this result is to choose (τ, H, B) and $\{P_h^n\}$ properly so that the bound obtained for that choice is *actually attained* by some estimator $\hat{\theta}_n$. Many examples of such choices are given in the remainder of the paper.

PROOF. It will be enough to give the proof for g bounded, uniformly continuous. Because of (2.11) and (5.8),

$$\delta_n(\theta_{nh} - \theta_0) \doteq -T^{-1} \circ \pi\xi_n(\theta_0, P_h^n) \doteq -T^{-1} \circ \pi[\Psi_n + V\tau h].$$

Therefore, the loss at time n is

$$(5.14) \qquad g(\delta_n|\tilde{\theta}_n - \theta_{nh}|) = g(|\tilde{\theta}_{n1} - T^{-1} \circ \pi \circ V\tau h|),$$

where $\tilde{\theta}_{n1}$ is just the "estimate"

$$\theta_{n1} = -\delta_n(\tilde{\theta}_n - \theta_0) - T^{-1} \circ \pi \Psi_n.$$

Let ϕ be the map of $B \to R^d$ given by $\phi = T^{-1} \circ \pi \circ V$. Define $\tau_1 : H \to R^d$ by $\tau_1 h = \phi \circ \tau h$. Let $B_1 = \tau_1 H$; let H_1 be any subspace of H such that $\tau_1 : H_1 \to B_1$ is one-to-one. Then (τ_1, H_1, B_1) is an abstract Wiener space, and, if Q_h^0 denotes its Gaussian shift experiment, then Q_0^0 is the distribution of $\phi(Z)$, where Z has distribution Q_0. Since the experiments $\{Q_h^0,\ h \in H_1\}$, $\{Q_h,\ h \in H_1\}$ are equivalent, $\{Q_h^n,\ h \in H_1\}$ converges to $\{Q_h^0,\ h \in H_1\}$. Therefore, by (5.14),

$$\begin{aligned}\inf_{\tilde{\theta}_n} \sup_{h \in N_c} \int g(\delta_n |\tilde{\theta}_n - \theta_{nh}|)\, dP_h^n \\ \geq \inf_{\tilde{\theta}_n} \sup_{h \in N_c \cap H_1} \int g(|\tilde{\theta}_n - \tau_1 h|)\, dP_h^n,\end{aligned} \tag{5.15}$$

where the second inf is over all estimates with values in B_1 (since $\tau_1 h \in B_1$, one can only do worse by taking estimates $\tilde{\theta}_n$ with values in $R^d - B_1$, by the projection theorem in R^d; the first inf in (5.15) is over all R^d-valued estimates). Since $l(x) \equiv g(|x|)$ is subconvex on B_1, the asymptotic minimax theorem of (4.10) shows the last expression in (5.15) is bigger than $\int g(|x|)\, dQ_0^0$, on letting $c \uparrow \infty$. Since Q_0^0 is the distribution of $\phi(Z)$, where Z has distribution Q_0, the result follows. Q.E.D.

(5.16) DEFINITION. A sequence of estimators $\hat{\theta}_n$ of $\{\theta_{nh},\ h \in H_0\}$ is called LAM if

$$\lim_c \lim_n \sup_{h \in N_c} \int g(\delta_n |\hat{\theta}_n - \theta_{nh}|)\, dP_h^n = Eg(Z).$$

A priori this definition depends on g and $\{N_c\}$; however, as a practical matter, in most applications the LAM property holds for a given $\hat{\theta}_n$ over a large variety of g, $\{N_c\}$.

(5.17) METHODOLOGICAL REMARK. In all applications of this paper, a selected sequence $\hat{\theta}_n$ of estimators is shown to be LAM by means of the following device. Pick c. Let h_n be an arbitrary sequence in N_c. One then shows that, under $P_{h_n}^n$, $\delta_n(\hat{\theta} - \theta_{nh_n})$ converges in distribution to Z, where Z was defined in (5.13). If, say, g is bounded, then this will indeed prove $\hat{\theta}_n$ LAM, since h_n can be chosen to satisfy

$$\sup_{h \in N_c} \int g(\delta_n |\hat{\theta}_n - \theta_{nh}|)\, dP_h^n = \int g(\delta_n |\hat{\theta} - \theta_{nh_n}|)\, dP_{h_n}^n$$

(or to come within n^{-1} of satisfying this relation).

(5.18) *Reformulation for $n^{1/2}$ scaling.* For convenience in later applications we write out the LAM result for the situation discussed in §3. Suppose (5.1)–(5.3) hold, and define θ_{nh} by (5.4). Replace hypothesis (5.7) by the assumption

$$n^{1/2} \xi_n(\theta_0, P_h^n) = \Psi_n + V \circ \tau h + o(1). \tag{5.19}$$

Assume further that the processes $\xi_n(\theta, P_h^n)$ satisfy the identifiability, differentiability, and boundedness assumptions of §3, with *nonsingular* derivative T independent of h. Under these conditions one has the following variant of the LAM result.

(5.20) THEOREM.

$$\lim_{c\uparrow\infty} \liminf_{n} \inf_{\tilde{\theta}_n} \sup_{h\in N_c} \int g(n^{1/2}|\tilde{\theta}_n - \theta_{nh}|)\, dP_h^n$$
$$\geq \int g(|T^{-1} \circ \pi \circ Vx|)\, Q_0(dx) = Eg(Z),$$

where $Z = T^{-1} \circ \pi \circ V \circ X$ and X is a B-valued random variable with distribution Q_0.

One naturally defines, in the present situation, a sequence of estimators $\hat{\theta}_n$ to be LAM if

$$\lim_{c} \lim_{n} \sup_{h\in N_c} \int g(n^{1/2}|\hat{\theta}_n - \theta_{nh}|)\, dP_h^n = Eg(Z).$$

Again, in all applications, a particular sequence $\hat{\theta}_n$ is proved to be LAM by showing that, under $P_{h_n}^n$, $n^{1/2}(\hat{\theta}_n - \theta_{nh_n})$ converges to Z whenever h_n is a sequence in some fixed N_c. Theorem (5.20) is proved by the method suggested in §3; see also Example (5.10).

VI. Weighted Cramer-von Mises estimation. The LAM results of this section were developed by Millar (1981). We therefore content ourselves in this section to showing how this particular example fits into the general framework of §§3, 5. For another investigation of this particular class of estimators, see Parr-de Wet and Parr-Schucany.

Let $X_1, \ldots, X_n$ be i.i.d. random variables. Let $\{P_\theta,\ \theta \in \Theta\}$ be a parametric family of measures on the line, where Θ is an open subset of R^d. Let

$$F(\theta; t) = P_\theta\{(-\infty, t]\} \tag{6.1}$$

be the cumulative distribution function of P_θ, and let

$$\hat{F}_n(t) = n^{-1} \sum_{1}^{n} I_{(-\infty,t]}(X_i) \tag{6.2}$$

be the empirical cdf.

Let μ be a measure on the line and $|\ |_\mu$ the $L^2(\mu)$ norm. For each θ, let $q_\theta(\cdot) \equiv q(\theta; \cdot)$ be a real function defined on the line; q is often called a *weight function.*

For real functions $a(\cdot)$ let

$$|a|_\theta^2 = \int a^2(s) q^2(\theta; s)\, \mu(ds) = |aq_\theta|_\mu^2. \tag{6.3}$$

The classic *Cramer-von Mises minimum distance estimate* (with weight function q) is defined by

$$\inf_{\theta} |\hat{F}_n - F_\theta|_\theta = |\hat{F}_n - F_{\hat{\theta}_n}|_{\hat{\theta}_n}. \tag{6.4}$$

For convenience, $\hat{\theta}_n$ will be referred to as CVMMDE. The choice $q(\theta; t) \equiv 1$ gives the "classic" minimum distance estimator mentioned in (1.1).

The goal of this section is to establish, in an appropriate framework, the local asymptotic optimality of the estimator $\hat{\theta}_n$.

The first step is to place the problem into the general framework of §§3, 5. To this end, fix a point $\theta_0 \in \Theta$; we shall establish a LAM property about "neighborhoods" of P_{θ_0}. Assume first that there is a sigma-finite measure m such that each P_θ has density $f(\theta,\cdot)$ w.r.t. m. Define H, the Hilbert space of real functions h, by

$$H = \left\{ h : \int h(s)^2 f(\theta_0; s)\, m(ds) < \infty, \right. \\ \left. \int h(s) f(\theta_0; s)\, m(ds) = 0,\ \text{support } h \subset \text{support } f(\theta_0,\cdot) \right\}. \tag{6.5}$$

Define, in addition,

$$\begin{aligned} &f(h;t) = f(\theta_0;t)[1+h(t)], \\ &H_0 = \{h \in H\colon f(hn^{-1/2};\cdot) \ge 0 \text{ for all large } n\}, \\ &F(h;\cdot) = cdf \text{ of } f(h;\cdot). \end{aligned} \tag{6.6}$$

Then H_0 is a dense convex subset of H and, if $h \in H_0$, $f(hn^{-1/2})$ is a probability density, at least for all large n. Define the product measure

$$P_h^n(dx) = \prod_1^n f(hn^{-1/2}; x_i)\, dx, \qquad x = (x_1, \dots, x_n). \tag{6.7}$$

Using the method suggested in (4.1), one easily sees that the experiments $\{P_h^n,\ h \in H_0\}$ converge to an experiment $\{Q_h,\ h \in H_0\}$. Indeed, let H be defined as above, and

$$\tau h(t) = \int^t h(s) f(\theta_0; s)\, m(ds).$$

If μ is finite (which we assume from now on; variants for sigma-finite μ are not difficult, but take a fair amount of space to discuss), then

$$\tau h \subset L^2(\mu). \tag{6.8}$$

Define the Hilbert space B by

$$B = \text{closure}\{\tau h\colon h \in H\}. \tag{6.9}$$

Then (τ, H, B) is an abstract Wiener space (cf. §4) and, if $\{Q_h,\ h \in H\}$ is its canonical normal shift family,

$$Q_0 \text{ is the distribution of } W_F^0, \tag{6.10}$$

where $W_F^0(t) = W^0(F(\theta_0;t))$, $t \in R^1$, and W^0 is the usual Brownian bridge on $[0,1]$. As usual, $Q_h(A) = Q_0(A - \tau h)$, $A =$ Borel set of $L^2(\mu)$. The convergence of $\{P_h^n,\ h \in H_0\}$ to $\{Q_h,\ h \in H_0\}$ was discussed in §4.

To state the basic LAM result, make the following assumptions and definitions. Assume that the maps

$$\theta \to q_\theta, \quad \theta \to F_\theta q_\theta, \quad \theta \to F_\theta \tag{6.11}$$

of Θ to $L^2(\mu)$ are Fréchet differentiable at θ_0 with derivatives $\eta_1, \eta_2, F'_{\theta_0}$, respectively. Denote by η the element of $L^2(\mu)$ given by

$$\eta = \eta_1 F_{\theta_0} - \eta_2 = -F'_{\theta_0} q_{\theta_0}. \tag{6.12}$$

The derivative η is *assumed nonsingular*. Assume further that

(6.13) if θ_n is any sequence in Θ such that $|F_{\theta_n} - F_{\theta_0}|_{\theta_n} \to 0$, then $\theta_n \to \theta_0$.

Define V, *multiplication operator* by

(6.14) $$Vz = q_{\theta_0} \times z, \qquad z \in L^2(\mu).$$

Let π be projection to the subspace of L^2 spanned by the components of η; define T as in (2.10). Define

(6.15) $$N_c = \{h \in H_0 \colon |h| \leq c\},$$

and let g, bounded, be defined as in §5. Define θ_{nh} by

(6.16) $$\inf_{\theta} |F(n^{-1/2}h) - F(\theta)|_\theta = |F(n^{-1/2}h) - F(\theta_{nh})|_{\theta_{nh}}$$

for $h \in H_0$. Recall $\hat{\theta}_n$, the "CVMMD" estimator of (6.4).

(6.17) THEOREM (LAM PROPERTY OF THE CVMMD ESTIMATE).

$$\begin{aligned}\lim_{c\uparrow\infty} \liminf_{n} \inf_{\tilde{\theta}_n} \sup_{h\in N_c} &\int g(n^{1/2}|\tilde{\theta}_n - \theta_{nh}|)\, dP_h^n \\ &= \int g(T^{-1} \circ \pi \circ V \circ z)\, Q_0(dz) \\ &= \lim_{c\uparrow\infty} \lim_{n} \sup_{h\in N_c} \int g(n^{1/2}|\hat{\theta}_n - \theta_{nh}|)\, dP_h^n.\end{aligned}$$

Here the inf is over all estimates $\tilde{\theta}_n$ of θ; see §7 for statistical interpretation of this result.

(6.18) REMARK. If the first equality is replaced by $\geq$, then the first half of the result remains valid if g is only increasing (boundedness of g is not necessary to establish the second expresion as the LAM lower bound).

(6.19) REMARK. The question here is really the estimation of a functional $\theta(\cdot)$ defined on distributions G by $G \to \theta(G)$,

(6.20) $$\inf_{\theta} |G - F_\theta|_\theta = |G - F_{\theta(G)}|_{\theta(G)}.$$

The functional is considered here only on the measures $F(n^{-1/2}h, dx)$ and $f(hn^{-1/2}; x)\, dx$, so $\theta(F(n^{-1/2}h)) = \theta_{nh}$; actually it can be considered on many more G—see §7 for some of them, and for reasons for such an extension. Part of the proof of (6.16) centers on showing that the map $F_{n^{-1/2}h} \to \theta(F_{n^{-1/2}h})$ is "differentiable" in the sense that $[\theta(F_{n^{-1/2}h}) - \theta(F_0)]n^{1/2}$ is, essentially, a linear function of h, as $n \to \infty$. For showing this, the basic result of §3 is the key. Evidently, there is an abstract LAM result that can be formulated for a wide class of "differentiable functions", not just those of the minimum distance type.

PROOF. Let us begin by proving the last equality in the theorem. For this, we shall use the method suggested in (5.17). Fix c and let h_n be an arbitrary sequence in N_c; we must show that $n^{1/2}(\hat{\theta}_n - \theta_{nh_n})$ converges, under P_{hn}^n, to $T^{-1} \circ \pi \circ V \circ Z$, where Z has distribution Q_0. Define processes

$$\xi_n^1(\theta, P_{hn}^n) = (F(n^{-1/2}h_n) - F_\theta)q_\theta, \quad \xi_n^2(\theta) = (\hat{F}_n - F_\theta)q_\theta.$$

Because

$$F(n^{-1/2}h) = F_{\theta_0} + \tau h n^{-1/2},$$

and because, under P_{hn}^n, $[\hat{F}_n - F(n^{-1/2}hn)]n^{1/2}$ converges in $L^2(\mu)$ to W_F (an application of the CLT for Hilbert space—see Parthasarathy (1967)), it is easy to see that ξ^1_h, ξ^2_h satisfy the *hypothesis of boundedness* as defined in §3 (under the sequence of measures P^{n}_{hn}). The hypothesis of convergence need NOT be satisfied here. Simple calculations show that (6.11), (6.13) imply, respectively, the hypotheses of *differentiability, identifiability* for *both* processes ξ^1_n, ξ^2_n; the derivative in each case, of course, is η. For this one needs to know, e.g., that $\hat{F}_n$, $F(n^{-1/2}h)$ both converge, under $P^n_{h_n}$, in $L^2(\mu)$ to F_{θ_0}. Therefore, the main result of §3 yields

$$\text{(6.21)} \qquad \begin{aligned} n^{1/2}(\hat{\theta}_n - \theta_0) &= -T^{-1} \circ \pi \circ \xi^2_n(\theta_0) + o(1), \\ n^{1/2}(\theta_{nh_n} - \theta_0) &= -T^{-1} \circ \pi \circ \xi^1_n(\theta_0) + o(1), \end{aligned}$$

and so

$$\text{(6.22)} \qquad n^{1/2}(\hat{\theta}_n - \theta_{nh_n}) = -T^{-1} \circ \pi \circ (\xi^2_n(\theta_0) - \xi^1_n(\theta_0)) + o(1).$$

Since

$$\text{(6.23)} \qquad \begin{aligned} \xi^2_n(\theta_0) - \xi^1_n(\theta_0) &= n^{1/2}[\hat{F}_n - F(n^{-1/2}h)]q_{\theta_0} \\ &\Rightarrow V \circ W_F \quad \text{under } P^n_{h_n}, \end{aligned}$$

the second equality of (6.17) follows from (6.18) and the boundedness of g.

Let us turn now to proving that the first expression in (6.16) is at least as big as the second expression there. To do this define

$$\text{(6.24)} \qquad \xi_n(\theta, P^n_h) = (F(n^{-1/2}h) - F_\theta)q_\theta.$$

Since $F(n^{-1/2}h) = F_{\theta_0} + n^{-1/2}\tau h$, we see that

$$\text{(6.25)} \qquad n^{1/2}\xi_n(\theta_0, P^n_h) = n^{1/2}\xi_n(\theta_0, P^n_0) + V \circ \tau h,$$

so that hypothesis (5.20) is satisfied. As in the first part of the proof, $\xi(\theta, P^n_h)$ satisfies the hypotheses of boundedness and differentiability—because of (6.11), (6.13). Therefore, the desired result is immediate from (5.21). Q.E.D.

VII. Comments on the statistical meaning of the result of §6. In this section we comment briefly on the statistical meaning of the setup advanced in §6; with minor changes, these comments carry over to all of the applications.

The observations $X_1, \ldots, X_n$ are i.i.d., but it is presumed that the common distribution F *may* be different from the distributions $\{P_\theta,\ \theta \in \Theta\}$ specified by the theoretical model. That is, the model does not *precisely* describe the data distribution, but it is presumed that it is not too unreasonable. There are a number of reasons why one might not have a parametric model that precisely describes the data. One possibility is that the observations have been subject to "data contamination" (roundoff errors, clerical errors, etc.). Another possibility is that the "true" model governing the X_i is so complicated (e.g., has inconveniently large dimension) that one would like to use a "simpler" model to fit the data as best one can, the idea being that one might not easily grasp the basic shape of the data if there are several million parameters.

Given that the data has a distribution that may differ from the model $\{P_\theta,\ \theta \in \Theta\}$, the first step is to specify what the possible data distributions might be. For *each* fixed $\theta_0 \in \Theta$, these are, in the model of §6, given by $\{F(n^{-1/2}h)\colon h \in N_c\}$, where c is an arbitrary fixed number. It is not difficult to see that this collection of data distributions can be augmented considerably. For example, the collection of data distributions just mentioned could be replaced by the collection of all cdf's F such that $\sup_t |F(t) - P_{\theta_0}(t)| \leq cn^{-1/2}$ (where $P_{\theta_0}(t)$ = cdf of P_{θ_0}), and the basic theorem of §6 will continue to hold. This allows an extremely large collection of possible data distributions; other minimum distance methods (cf. §§13, 12) will admit only relatively *small* collections of data distributions, a feature that sometimes is a severe disadvantage.

Having specificed the model $\{P_\theta,\ \theta \in \Theta\}$ and the possible data distributions, the task is to 'fit the model to the data as best one can'. As discussed in §1, there are a great many possibilities, depending on what the statistical goals are. The criterion for 'best fit' adopted in §6 is given by (6.16), (6.20); for this to be acceptable it is more or less tacit that the statistical goals include the estimation of a number of probabilities; if one wanted to estimate moments as well, a somewhat different notion of best fit should be adopted (see §13, Example 4 for one possibility); and if one wanted quantiles estimated too, then probably yet another framework should be set up (cf. §13, Example 6(c)). In view of the last sentence of the preceding paragraph, the notion of best fit to be adopted *will also have* to take into account the kind of *data* distributions you believe are present: if G is a possible data distribution you evidently at least have to be able to say what the 'distance' between G and $\{P_\theta\}$ is. If the class of data distributions is believed to be rather broad, the approach of §6 is appropriate; if you are sure the possible data distributions form a fairly small class, then you could use some of the other methods in §13.

Having specified the possible data distributions and an appropriate notion of 'best fit' (in the case under discussion it is specified by (6.20)), we wish to find the point θ that 'fits' P_θ as close as possible to the actual (unknown) data distribution G. That is, we wish to *estimate* $\theta(G)$, defined in (6.19). The result (6.17) is an optimality theorem elaborated in just such a framework. Such LAM results have a long history, going back to classical parametric estimation; discussion of their statistical importance should be unnecessary at this point in time (see Hajek (1972), Le Cam (1972), Beran (1981), Millar (1981) for original or recent developments). Let it suffice to say, the intent of LAM theorems is to ensure at least a desirable stability of estimators in the convergence to their asymptotic limit.

Questions can, of course, be raised about the choice of the particular framework of §6 (and for all of the other applications). Fortunately, the setup of §6 has been heavily studied. A certain amount of confidence in the excellence of the procedures advanced there derives from the fact that they actually work well in practice. See Parr-Schucany for some empirical studies; see Millar (1981) for arguments (based on (2.12), (2.16)) showing their asymptotic equivalence to other procedures that have been advanced. That these procedures should behave decently is demonstrated theoretically in the proof of (6.17), where the crux of the matter was to show convergence to the asymptotic limit, that is uniform over the *large* neighborhoods N_c. Finally, if the possible data distributions are truly broad, as assumed in §6, the LAM framework nicely rules out certain 'unstable' estimators. For example, if P_θ were $N(\theta, 1)$ one could use $\overline{X}$, the sample mean, to estimate the θ that 'fits best'; it is easy to see this will not be LAM, since it is always possible to find neighboring data distributions G, $|G - P_{\theta_0}| \leq n^{-1/2}$, G with *no* mean; this will force $\overline{X}$ to be extremely unstable asymptotically (see Millar (1981) for the precise calculation). Evidently, the entire framework of §6 could be cast in terms of robustness; but this was done already in Millar (1981); robustness terminology can also be appled to some of the other developments; we leave this to the reader.

VIII. Minimum χ^2 estimates. The estimation method of this section is a simple variant of those of §6; it is worth singling out because minimum χ^2 methods have enormous importance in applied statistics.

Let $\{P_\theta,\ \theta \in \Theta\}$ be a family of probabilities concentrated on the integers $1, \ldots, d$ and indexed by an open subset $\Theta \subset R^s$, $s \le d$. Denote by

$$\begin{aligned} &F_\theta(t) = F(\theta; t) \text{ the cdf of } P_\theta, \\ &f(\theta; i) = f_i(\theta) = \text{mass assigned to } \{i\} \text{ by } P_\theta, \\ &f(\theta) = f(\theta; \cdot) \text{ the vector } (f(\theta; 1), \ldots, f(\theta; d)) \in R^d. \end{aligned} \tag{8.1}$$

If G is any probability on the integers $1, \ldots, d$ with $g_i = G(\{i\})$, define the *minimum χ^2 functional* $\theta(G)$ by

$$\inf_\theta \sum \frac{[g_i - f_i(\theta)]^2}{f_c(\theta)} = \sum \frac{[g_i - f_i(\theta(G))]^2}{f_i(\theta(G))}. \tag{8.2}$$

As in other examples, some tedium can be avoided if it is assumed that this inf is achieved at some point; otherwise the usual trivial device needs to be employed (take *any* $\hat{\theta}(G)$ that comes within n^{-2} of achieving the inf). Define, for any $x = (x_1, \ldots, x_d) \in R^d$,

$$|x|_\theta^2 = \sum \frac{x_i^2}{f_i(\theta)}. \tag{8.3}$$

Then if g is the vector $(g_1, \ldots, g_d)$, $\theta(G)$ satisfies

$$\inf_\theta |g - f_\theta|_\theta = |g - f_{\theta(G)}|_{\theta(G)}, \tag{8.4}$$

a formulation which emphasizes the similarity of $\min \chi^2$ methods to minimum distance Cramer-von Mises methods of §6.

Assume that $X_1, \ldots, X_n$ are i.i.d. random variables with values in $(1, \ldots, d)$. Let $\hat{F}_n$ be the empirical cdf and let

$$\hat{f}(i) = \hat{F}_n(\{i\}), \tag{8.5}$$

the mass assigned to $\{i\}$ by $\hat{F}_n$. The $\min \chi^2$ *estimate of θ* is

$$\theta(\hat{F}_n). \tag{8.6}$$

The goal of this section is to show that $\theta(\hat{F}_n)$ is asymptotically normal and LAM; normality is well known.

To describe these results, fix $\theta_0 \in \Theta$. Define

$$H = \left\{ h \in R^d \colon \int h\, dF_{\theta_0} = \sum h_i f_c(\theta_0) = 0 \right\}. \tag{8.7}$$

For $h \in H$, define probabilities $F(h)$ on $\{1, \ldots, d\}$ by specifying that the mass assigned to $\{i\}$ by $F(h)$ shall be $f(h; i) = f_i(h)$, defined by

$$f(h; i) = f(\theta_0; i)[1 + h_i] \tag{8.8}$$

if $h = (h_1, \dots, h_d)$. See §6 for an extremely close analogue. Let

(8.9) P_h^n, $h \in H$, be the n-fold product of $F(n^{-1/2}h)$.

For each $c > 0$ define

$$N_c = \{h \in H\colon |h|_0 \le c\}, \tag{8.10}$$

where $|\ |_0$ is the norm of R^d.

Next let us introduce hypotheses on the family $\{F_\theta,\ \theta \in \Theta\}$. Assume:

(8.11)
(a) The map $\theta \to f(\theta)$ of Θ to R^d is differentiable.
(b) The derivative is bounded in a neighborhood of θ_0.
(c) For each i, $f_i(\theta)$ is strictly positive near θ_0.

Define vectors $\eta_1, \dots, \eta_s$, $\eta_i \in R^d$, such that the jth component of η_i is

$$(\eta_i)_j = [f^{1/2}(\theta; j)]^{-1}[(\partial/\partial\theta_i) f(\theta; j)_{\theta=\theta_0}], \qquad 1 \le j \le d.$$

Assume the nonsingularity hypothesis

(8.12) $\eta_1, \dots, \eta_s$ are linearly independent.

Define:

(8.13)
π: projection in R^d to span η_i.
T: map of Θ to span η_i by $T\theta = \langle\theta, \eta\rangle$,

where $\langle\theta, \eta\rangle = \sum \theta_i \eta_i$ when $\theta = (\theta_1, \dots, \theta_s)$. Since η is nonsingular, T^{-1} exists. Assume, in addition,

(8.14) if $|f(\theta_n) - f(\theta_0)|_{\theta_n} \to 0$, then $\theta_n \to \theta_0$.

Define the multiplication operator V on R^d by

$$(Vx)_i = x_i / f^{1/2}(\theta_0; i), \qquad x = (x_i, \dots, x_d). \tag{8.15}$$

Let Q_0 be the normal distribution on R^d with mean 0 and covariance matrix Σ, where the (i-j)th element of Σ is

$$\Sigma_{ij} = \begin{cases} -f(\theta_0; i) f(\theta_0; j), & i \ne j, \\ f(\theta_0; i)[1 - f(\theta_0; i)], & i = j. \end{cases} \tag{8.16}$$

Finally, let l be a function on R^d of the form

$$l(x) = g(|x|_0) \tag{8.17}$$

where g is a bounded increasing function on $[0, \infty)$. Then the following result establishes the LAM character of the minimum χ^2 estimates.

(8.18) THEOREM.

$$\begin{aligned} &\lim_{c\uparrow\infty} \liminf_{n} \inf_{\tilde\theta_n} \sup_{n \in N_c} \int l(n^{1/2}(\tilde\theta_n - \theta_{nh}))\, dP_h^n \\ &\quad = \int l(T^{-1} \circ \pi \circ Vx)\, Q_0(dx) \\ &\quad = \lim_{c\uparrow\infty} \lim_{n} \sup_{h \in N_c} \int l(n^{1/2}(\hat\theta_n - \theta_{nh}))\, dP_h^n, \end{aligned}$$

where $\theta_{nh} = \theta(F(n^{-1/2}h))$, $\hat{\theta}_n = \theta(\hat{F}_n)$ *and the inf is taken over all estimates* $\tilde{\theta}_n$ *of* θ.

Industrious readers may use (3.8) to write out the covariance matrix of $T^{-1} \circ \pi \circ V$.

PROOF. Let us first establish the second equality. For this it is enough (cf. (5.17)) to show that if h_n is any sequence in N_c, then, under P_h^n, $n^{1/2}(\hat{\theta}_n - \theta_{nh_n}) \Rightarrow T^{-1} \circ \pi \circ VZ$, where Z has distribution Q_0. Let ξ_n^1 be the random vector with ith component $[\hat{f}_n(i) - f_i(\theta)]/f_i^{1/2}(\theta)$. Since, under P_h^n,

$$n^{1/2}[\hat{f}_n - f(n^{-1/2}h_n)] \Rightarrow Z, \tag{8.19}$$

it is clear that ξ_n^1 satisfies the hypothesis of *boundedness* given in §3. Hypotheses (8.11), (8.12) and (8.14) ensure differentiability (with nonsingular derivative η) and identifiability respectively (as these are defined in §3). Therefore, by (3.6),

$$n^{1/2}(\hat{\theta}_n - \theta_0) = -T^{-1} \circ \pi \circ \xi_n^1(\theta_0) + o(1). \tag{8.20}$$

Similarly, one shows

$$n^{1/2}(\theta_{nh_n} - \theta_0) = -T^{-1} \circ \pi \circ \xi_n^2(\theta_0) + o(1), \tag{8.21}$$

where $\xi_n^2(\theta)$ is the vector whose ith component is $[f(n^{-1/2}h_n; i) - f(\theta; i)]/f^{1/2}(\theta; i)$. Combining (8.20), (8.21) and using the definition of V yields

$$n^{1/2}(\hat{\theta}_n - \theta_{nh_n}) = -T^{-1} \circ \pi \circ V \circ (\hat{f}_n - f(n^{-1/2}h_n)) + o(1), \tag{8.22}$$

which gives the desired conclusion, by (8.19).

To finish the proof of the theorem, it is necessary now to show only that the first expression is at least as big as the second, and, as in other cases, one does not need to assume that g is bounded in order to carry this out. To establish this LAM lower bound, we apply the result of (5.21). For this, define vector processes $\xi_n(\theta, P_h^n)$ with values in $\{1, \ldots, d\}$ by specifying the ith component of $\xi_n(\theta, P_h^n)$ to be

$$[f(n^{-1/2}h; i) - f(\theta; i)]/f^{1/2}(\theta; i). \tag{8.23}$$

Define τ_0, the mapping on H, by

$$(\tau_0 h)(t) = \int^t h(s) F(\theta_0; ds),$$

where $h(s)$ is the step function on $[0, d]$ obtained from $(h_1, \ldots, h_d)$. Let Q_{00} be the distribution of $W^0(F(\theta_0; t))$ and Q_{0h} its shifts by τ_0. Then, as in §6, the experiments $\{P_h^n,\ h \in H\}$ converge to $\{Q_{0h}\}$. The measure Q_{00} may be identified with the distribution of $(W^0(F(\theta_0; 1)), \ldots, W^0(F(\theta_0; d)))$. Define a mapping τ_1 on R^d by $\tau_1 x = (x_1, x_2 - x_1, \ldots, x_d - x_{d-1})$. Let Q_0 be the image of Q_{00} under τ_1; of course, Q_0 is the same as the measure defined in (8.16). Define $\tau = \tau_1 \circ \tau$, a mapping on H, and let Q_h be the shift of Q_0 by τ. Then since τ_1 is one-to-one, $\{Q_n\}$ is equivalent to $\{Q_{0h}\}$, so $\{P_h^n\}$ converges to $\{Q_h\}$. Moreover,

$$n^{1/2}\xi_n(\theta_0, P_h^n) = V \circ \tau \circ h$$

since $(\tau h)_i = h_i f(\theta_0; i)$. As in the first part of the proof, the processes $\xi_n(\theta, P_h^n)$ satisfy the hypotheses of convergence, differentiability, and identifiability. The LAM lower bound is now immediate from (5.12). Q.E.D.

IX. Estimates based on quantile functions. Let G be a probability cdf on the line. Let $\mathcal{Q}$ be the *quantile function* of G: the function defined on $s \in [0,1]$ by

$$\mathcal{Q}(G)(s) = \inf\{t\colon G(t) \geq s\}. \tag{9.1}$$

That is, $\mathcal{Q}(G)$ is the usual right continuous inverse of the increasing function $t \to G(t)$. We shall often write $G^{-1}(s)$ for $\mathcal{Q}(G)(s)$. Let $X_1, \ldots, X_n$ be real random variables; if $\hat{F}_n$ is the empirical cdf then

$$\mathcal{Q}(\hat{F}_n) \text{ is the empirical quantile function.} \tag{9.2}$$

Let now $\{P_\theta,\ \theta \in \Theta\}$ be a fixed family of probabilities on the line indexed by an open subset Θ of R^d. Let

$$F(\theta; t) = F_\theta(t) = \text{cdf of } P_\theta. \tag{9.3}$$

The quantile minimum distance functional $\theta(G)$ is defined on cdf's G by

$$\inf_\theta |\mathcal{Q}(G) - \mathcal{Q}(F_\theta)| = |\mathcal{Q}(G) - \mathcal{Q}(F_{\theta(G)})|, \tag{9.4}$$

where $|\ |$ denotes the norm of an appropriate Hilbert space wherein the functions G^{-1}, F_θ^{-1} are alleged to lie. The *quantile minimum distance estimate* of θ is

$$\theta(\hat{F}_n) \equiv \hat{\theta}_n, \tag{9.5}$$

to be denoted QMDE occasionally.

The purpose of this section is to describe the LAM optimality properties of the QMDEs of θ. As in other sections, it is to be understood here that the statistical goal is to obtain the "best fit" of the model to the data, where "best fit" now means matching up the quantile functions as best as one can—a not unreasonable statistical purpose for many investigations.

We shall assume $X_1, \ldots, X_n$ i.i.d. and, under suitable hypotheses, show $\mathcal{Q}(\hat{F}_n)$ to be asymptotically normal and LAM. Asymptotic normality has been discussed by La Riccia, using a rather less abstract point of view; the LAM aspect appears here for the first time.

Let $0 < \alpha < \beta < 1$ and let μ be a finite measure on $[\alpha, \beta]$. To avoid difficulties with the tails of G^{-1}, we shall look at G^{-1} only on the interval $[\alpha, \beta]$, and, with this convention, shall regard G^{-1} as an element of $L^2(\mu)$. In particular, the minimum distance functional $\theta(G)$ is to be computed using $|\ |$, the norm of $L^2(\mu)$. Obviously more technically involved set-ups are possible: one can take μ on $(0,1)$, but then additional hypotheses are necessary so that $G^{-1} \in L^2(\mu)$ for the G arising below; in particular, $\mathcal{Q}(\hat{F}_n) \notin L^2(\mu)$ for general μ unless certain conventions are adopted. This superficially greater generality involves no truly new insight, while making the situation relatively messy; therefore, we content ourselves with our present framework. For the parametric family $\{P_\theta\}$, assume:

(9.6) (a) support $F^{-1} \supset [\alpha, \beta]$.

(b) P_θ has density f_θ with respect to Lebesgue measure, f_θ continuous and strictly positive on the interior of its support.

(c) If $F_{\theta_n}^{-1} - F_{\theta_0}^{-1}$ converges to 0 in $L^2(\mu)$, then $\theta_n \to \theta_0$.

(d) $F^{-1}(\theta) - F^{-1}(\theta_0) = \langle \theta - \theta_0, \eta \rangle + o(1)$ in $L^2(\mu)$, where $\eta = (\eta_1, \ldots, \eta_d)$, $\{n_i\}$ a linearly independent set of elements of $L^2(\mu)$.

Under a bit more regularity, and with some effort, hypotheses (c), (d) can be put directly in terms of the cdf's F_θ; in some examples (e.g., location model) there is clearly no advantage to doing so. Define H to be the Hilbert space of real functions h on $[0,1]$, such that $\int h^2(s)\,ds < \infty$, $\int h(s)\,ds = 0$. Let $F_h = F(h)$ be the cdf with density

$$dF_h(t) = f(\theta_0;t)[1 + h(F(\theta_0;t))].$$

Define τ on H by $(\tau h)(t) = \int^{F(\theta_0;t)} h(s)\,ds$, so

$$F_h = F_{\theta_0} + \tau h. \tag{9.7}$$

Let B = closure of τH under sup norm: so B is a subcollection of continuous functions of the form $x(F(\theta_0;t))$, $x \in C[0,1]$. Then (τ, H, B) is an abstract Wiener space. Let $\{Q_h\}$ be its Gaussian shifts experiment; then Q_0 is the distribution of $\{W^0(F(\theta_0;t))\}$.

Let P_h^n be the n-fold product of $F(n^{-1/2}h)$; then $\{P_h^n\}$ converges to $\{Q_h\}$. Let $B_2 = L^2(\mu)$ and define the map $V\colon B \to B_2$ by

$$(Vx)(t) = [x(F^{-1}(\theta_0;t))/f(F_{\theta_0}^{-1}(t))]I_{[\alpha,\beta]}(t),$$

where $f(s) = f(\theta_0;s)$. Define π, projection in B_2 to span $\{\eta_i\}$, and define T as usual (cf. (2.10)).

Let $N_c = \{h \in H\colon |h| \le c$, h has compact support in $(0,1)$, infinitely differentiable, with bounded derivatives$\}$.

(9.8) THEOREM.

$$\begin{aligned}\lim_{c\uparrow\infty}\liminf_{n}\inf_{\tilde\theta_n}\sup_{h\in N_c} &\int g(n^{1/2}|\tilde\theta_n - \theta_{nh}|)\,dP_h^n\\ &= \int g(|T^{-1}\circ\pi\circ Vx|)\,Q_0(dx)\\ &= \lim_{c\uparrow\infty}\lim_{n}\sup_{h\in N_c}\int g(n^{1/2}|\hat\theta_n - \theta_{nh}|)\,dP_h^n,\end{aligned}$$

where $\hat\theta_n$ is the QMDE.

PROOF. Define

$$\xi_n(\theta, P_h^n)(t) = [F^{-1}(n^{-1/2}h;t) - F^{-1}(\theta;t)]I_{[\alpha,\beta]}(t).$$

If $h \in N_c$ then, in B_2,

$$\sqrt{n}\,\xi_n(\theta_0, P_h^n) = V\circ\tau h + o(1) \tag{9.9}$$

using $F(n^{-1/2}h) = F(\theta_0) + n^{-1/2}\tau h$. The calculation (9.9) has long been familiar to students of robustness with varying degrees of rigor (cf. Huber (1977) and Serfling (1980)). Hypothesis (9.6) guarantees that identifiability and differentiability hold for $\xi_n(\theta, P_h^n)$ whenever $h \in N_c$. It is now immediate from the basic result of §5 that the first expression in (9.8) is at least as big as the second. To see why the QMDE is optimal (i.e., to prove the second inequality and hence the theorem), let h_n be a sequence in N_c; we must show that under $P_{h_n}^n$, $n^{1/2}(\hat\theta_n - \theta_{nh_n})$ converges to $T^{-1}\circ\pi\circ V\circ Z$, where Z has the distribution of $W^0(F(\theta_0;t))$. Relation (9.9) suggests that if $\xi_{1n}(\theta) = \hat F_n^{-1} - F^{-1}(\theta)$, then

$$n^{1/2}\xi_{1n}(\theta_0) = V(n^{1/2}(\hat F_n - F(\theta_0))) + o(1)$$

under $P^n_{h_n}$. Arguments of Shorack can be used to make this rigorous. The process $\xi_{1n}(\theta)$ therefore satisfies the hypotheses of boundedness (under $P^n_{h_n}$); it satisfies identifiability and differentiability by (9.6). Therefore

$$n^{1/2}(\hat{\theta}_n - \theta_0) = -T^{-1} \circ \pi \circ V(n^{1/2}(\hat{F}_n - F_{\theta_0})) + o(1)$$

by the basic result of §3. Similarly, if $\xi_{2n}(\theta) = F^{-1}(n^{-1/2}h_n) - F^{-1}(\theta)$, then

$$n^{1/2}(\theta_{nh} - \theta_0) = -T^{-1} \circ \pi \circ V(n^{1/2}[F(n^{1/2}h_n) - F(\theta_0)]),$$

so together,

$$n^{1/2}(\hat{\theta}_n - \theta_0) = -T^{-1} \circ \pi \circ V(n^{1/2}(\hat{F}_n - F(n^{-1/2}h))) + o(1).$$

Since, under $P^n_{h_n}$, $n^{1/2}[\hat{F}_n - F(n^{-1/2}h_n)]$ converges to $W^0(F(\theta_0))$, the result is proved. Q.E.D.

X. Simple regression. Let F be a fixed *known* cdf on the line. For each n, let $c_{n1}, \ldots, c_{nn}$ be known real numbers. Let $X_{n1}, \ldots, X_{nn}$ be real random variables. The *simple regression* model asserts that, for some θ,

(10.1) $$\{X_{ni} - c_{ni}\theta,\ 1 \leq i \leq n\} \text{ is i.i.d. } F.$$

The statistical problem is to "estimate θ".

Let μ be a measure on the line; set $B = L^2(\mu)$. Let $d_{n1}, \ldots, d_{nn}$ be a sequence of numbers such that

(10.2) $$\max_i |d_{ni}| \to 0, \qquad \sum d^2_{ni} = 1.$$

Define the process

(10.3) $$\xi_n(\theta)(t) = \sum d_{ni}[I\{X_{ni} \leq t + c_{ni}\theta\} - F(t)],$$

the "normalized weighted empirical cdf". $\xi_n(\theta)$ is a random element that we suppose belongs to $L^2(\mu)$. The minimum distance estimate of θ is $\hat{\theta}_n$, defined by

(10.4) $$\inf_\theta |\xi_n(\theta)|_\mu = |\xi_n(\hat{\theta}_n)|_\mu.$$

Under various hypotheses, Koul-de Wet proved that such an estimate was asymptotically normal. We explain in this section how this asymptotic normality follows from our general structural theorems. In addition, we prove a LAM property which is new.

To state the LAM property let

(10.5) $$f \text{ be the density of } F$$

with respect to Lebesgue measure. We suppose $f \in L^2(\mu)$. Define a Hilbert space $H = \{h \in L^2(\mu)\colon \int hf = 0\}$. Construct probabilities $F(h; dx)$ by specifying its density $f(h; x)$ to be

(10.6) $$f(h; x) = f(x)[1 + h(x)].$$

Define P^n_h, a product measure, so that under P^n_h, $X_{ni} - c_{ni}\theta_0$, $1 \leq i \leq n$, are independent with $X_{ni} - c_{ni}\theta_0$ distributed as $F(d_{ni}h)$. Define $(\tau h)(t) = \int^t h(s)f(s)\,ds$, and note that

(10.7) $$F(d_{ni}h) = F + d_{ni}\tau h.$$

Assume μ finite, for convenience (so $\tau h \in L^2(\mu)$); define B to be the closure of τH in $L^2(\mu)$. Then (τ, H, B) is an abstract Wiener space; if $\{Q_h\}$ is its Gaussian shift experiment, then Q_0 is the distribution of $W^0(F)$. Let $N_c = \{h \in H\colon |h| \leq c\}$. Define θ_{nh} by

$$(10.8) \qquad \xi_n(\theta; P_h^n)(t) = \sum d_{ni}[F(d_{ni}h; t) - F(t - c_{ni}(\theta - \theta_0))],$$
$$\inf_\theta |\xi_n(\theta, P_h^n)|_\mu = |\xi_n(\theta_{nh})|_\mu.$$

Let Z be a real Gaussian random variable,

$$(10.9) \qquad Z \sim N(0, \sigma^2),$$

$$(10.10) \qquad \sigma^2 = \left(\frac{1}{|f|_\mu}\right)^4 \iint [F(s \wedge t) - F(s)F(t)] f(s) f(t)\, \mu(ds)\, \mu(dt).$$

The following theorem holds under the conditions given in Koul-de Wet; these conditions, which include restrictions on the c_{ni}, d_{ni}, are quite lengthy to state.

(10.11) THEOREM.

$$\lim_c \liminf_n \inf_{\tilde{\theta}_n} \sup_{h \in N_c} \int g(\delta_n |\tilde{\theta}_n - \theta_{nh}|)\, dP_h^n = Eg(|Z|)$$
$$= \lim_c \lim_n \sup_{h \in N_c} \int g(\delta_n |\hat{\theta}_n - \theta_{nh}|)\, dP_h^n,$$

where $\delta_n = \Sigma\, d_{ni} c_{ni}$, and $\hat{\theta}_n$ was defined in (10.4).

PROOF. The proof will be sketched only; a number of the calculations can be found in Koul-de Wet. We begin by showing that the first expression in the theorem is bigger than the second. Because of (10.2) and (10.8), $\xi_n(\theta_0, P_h^n) = \tau h$, so condition (5.7) is satisfied as well as the condition of boundedness (2.4). Moreover, under mild conditions,

$$\xi_n(\theta, P_h^n) - \xi_n(\theta_0, P_h^n) = -\sum d_{ni}[F(t - c_{ni}(\theta - \theta_0)) - F(t)]$$
$$\doteq \sum c_{ni}\, d_{ni}(\theta - \theta_0) f(t)$$
$$= \delta_n(\theta - \theta_0) f(t) \quad \text{in } L^2(\mu),$$

so ξ_n is differentiable with derivative $\eta = f$ as defined in (3.2). The hypothesis of identifiability is more delicate and depends on relations between c_{ni}, d_{ni}, as well as some smoothness in F. It can be checked under the conditions of Koul-de Wet. Therefore, the processes $\xi_n(\theta, P_h^n)$ satisfy the condition of §5, so that the first expression in the statement of the theorem is at least as big as

$$(10.12) \qquad \int g(|T^{-1} \circ \pi x|)\, Q_0(dx),$$

where π is projection to span η in $L^2(\mu)$, and T is the usual operator (2.10). But here,

$$T^{-1} \circ \pi x = \langle x, f \rangle_\mu / |f|_\mu^2, \qquad x \in L^2(\mu),$$

by (2.12), and Q_0 is the distribution of $W^0 \circ F$. Therefore, (10.12) is equal to $Eg(|Z|)$, where $Z = \langle W_F, f \rangle_\mu / |f|_\mu^2$. By an elementary calculation using the form of the covariance function of W_F and the form of the $L^2(\mu)$ inner product, it is easy to see that the Z just given has the distribution (10.10). This establishes the LAM lower bound.

To show that $\hat{\theta}_n$, defined by (10.4), is LAM, let h_n be a sequence in N_c. It is necessary to show only that $\delta_n(\hat{\theta}_n - \theta_{nh_n})$ converges to Z under $P^n_{h_n}$. This was shown by Koul-de Wet for $\delta_n(\hat{\theta}_n - \theta_n)$ under P^n_0; we content ourselves here with showing how to apply the abstract approach of §2.

Under $P^n_{h_n}$, one first shows that the process $\xi_n(\theta)$, defined in (10.3), satisfies the conditions of boundedness, identifiability, and differentiability, with derivative $\delta_n f$ (same as for $\xi_n(\theta, P^n_h)$ introduced in (10.9)). This undertaking involves a fair amount of work. To see that differentiability is at least plausible, write

$$\xi_n(\theta_0 + \theta) - \xi_n(\theta_0) = \sum d_{ni} I\{c_{ni}\theta_0 + t \leq X_{ni} \leq c_{ni}(\theta + \theta_0) + t\},$$

which is possible if $c_{ni}\theta \geq 0$; it is then plausible that this last expression is approximable, in probability, in $L^2(\mu)$, by $\Sigma c_{ni} d_{ni} f(t)\theta$. The hypothesis of boundedness, under $P^n_{h_n}$ can be checked, e.g., by using the CLT in Hilbert space to show that, under $P^n_{h_n}$,

$$\xi^*_n(\theta) \equiv \sum d_{ni}[I\{X_{ni} \leq t + c_{ni}\theta_0\} - F(n^{-1/2}h_n)]$$

converges in $L^2(\mu)$ to $W^0(F)$; the treatment is a minor variant of the usual arguments involving the empirical cdf. Identifiability is more involved; see Koul-de Wet for applicable arguments. Assuming this done, the theory of §2 shows

$$\delta_n(\hat{\theta}_n - \theta_0) = -T^{-1} \circ \pi \circ \xi_n(\theta_0) + o(1).$$

As in the first part of the proof the processes $\xi_n(\theta, P^n_{h_n})$ satisfy the basic three hypotheses, so

$$\delta_n(\theta_{nh} - \theta_0) = -T^{-1} \circ \pi \circ \xi_n(\theta_0, P^n_{h_n}) + o(1).$$

Therefore

$$\delta_n(\hat{\theta}_n - \theta_{nh}) = -T^{-1} \circ \pi \circ \xi^*_n(\theta) \Rightarrow -T^{-1} \circ \pi \circ W_F,$$

whence the result. Q.E.D.

(10.13) REMARK. It may be worthwhile at this point to make more explicit the sense in which the present "estimate of θ" is optimal. The model asserts that the data $X_{n1}, \ldots, X_{nm}$ are independent, with distributions given by the vector $V_n(\theta) = F(t - c_{ni}\theta), \ldots, F(t - c_{nn}\theta)$ for some θ. On the other hand, due to "data contamination", it may happen that the data are independent, with distributions governed by a sequence of cdf's $G_n = (G_{n1}, \ldots, G_{nn})$. In these circumstances one announces the statistical goal to be that of finding the point θ such that the vector of cdf's, $V_n(\theta)$, "best fits the data", i.e., such that $V_n(\theta)$ is as "close as possible" to G_n. Many notions of distance could be chosen here; the distance used in this section was the $L^2(\mu)$ distance between $\Sigma d_{ni} G_{ni}$ and $\Sigma d_{ni} F(\cdot - c_{ni}\theta)$. When the d_{ni} have been chosen, the main result says that the estimate $\hat{\theta}_n$ of (10.4) does the best job under the situation where there is rather severe local contamination (i.e., quite a few G_n are allowed), and *when the goal is the one of "best fit" just mentioned.* It seems quite an irrelevant undertaking, in the present context, to compare the asymptotic variances of the present estimate with others that may be invented.

One could choose the d_{ni} (subject to (10.2)) so that δ_n is as large as possible; this would ensure that the estimate $\hat{\theta}_n$ has the "fastest convergence" to θ_{nh}. This choice is $d_{ni} = c_{ni}/(\Sigma c_{ni}^2)^{1/2}$ by the Schwarz inequality. On the other hand, the selection of d_{ni} determines what you mean by "best fit" described in the preceding paragraph; and this newly proposed choice of d_{ni} may or may not embody a reasonable notion of "best fit". It is not unthinkable that practitioners would prefer $d_{ni} = n^{-1/2}$, $1 \leq i \leq n$, since this yields a readily grasped notion of closeness. In other words, the choice of d_{ni} should depend on what you think the best notion of distance is for fitting the model to the data. If you, for example, choose the d_{ni} unequal, this means you want certain of the G_i to be better "fitted" by the model than others; there may be good reason for this in some cases, but such a reason would not depend on the size of δ_n.

To put it another way: when you change the d_{ni}, you change the notion of *which* θ you are trying to estimate. You should decide at the beginning, on the basis of your statistical goals, *which* θ you want to estimate. For this it is irrelevant that, if someone suggests estimating a *different* θ, then some estimator for *that* θ has, e.g., a faster convergence rate.

XI. Estimates based on spectral functions. Let $X_1, \ldots, X_n$ be a real stationary Gaussian sequence with mean 0. Assume that the spectral function is of the form $F(\theta; t)$, $-\pi \leq t \leq \pi$, for some $\theta \in \Theta$ an open subset of R^d; then

$$E_\theta X_k X_{k+j} = \int_{-\pi}^{\pi} e^{ijt} F(\theta; dt). \tag{11.1}$$

Assume $F(\theta; 0) = 0$; this reduction involves only a renormalization. Define, for any spectral function G ($G(o) = 0$), the minimum distance functional $\theta(G)$ by

$$\inf_\theta |G - F(\theta)|_\mu = |G - F(\theta(G))|_\mu, \tag{11.2}$$

where μ is a finite measure on $[0, \pi]$ and $|\ |_\mu$ is the norm of $L^2(\mu)$. Define

$$\hat{F}_n(t) = (2\pi n)^{-1} \int_0^t \left| \sum_{k=1}^{n} X_k e^{-iyk} \right|^2 dy. \tag{11.3}$$

Then $\hat{F}_n$ is a random element of $L^2(\mu)$ and is a well-known spectral function estimate (closely analogous to the empirical cdf as an estimate of a cdf, discussed in §6). Define $\hat{\theta}_n$, estimate of θ, by

$$\hat{\theta}_n = \theta(\hat{F}_n). \tag{11.4}$$

The main result of this section is that the estimator $\hat{\theta}_n$ is asymptotically normal and LAM; these results appear to be new. The development is based on the general theory of §§3, 5 and illustrates the fact that "independent observations" have nothing to do with the basic framework.

To state these results, fix θ_0. Assume that

(11.5) (a) $F(\theta_0; \cdot)$ has density $f(\theta_0; t)$ on $[0, \pi]$ with respect to Lebesgue measure, and $\int_0^\pi f^2(t)\, dt < \infty$.

(b) If $F(\theta_n) \to F(\theta_0)$ in $L^2(\mu)$, then $\theta_n \to \theta_0$.

(c) There exists $\eta = (\eta_1, \ldots, \eta_d)$, η_i linearly independent elements of $L^2(\mu)$ such that $F(\theta) - F(\theta_0) = \langle \theta, \eta \rangle + o(|\theta - \theta_0|)$.

Let H be the Hilbert space of all real functions h on $[0, \pi]$ such that $\int_0^\pi h^2 f^2 < \infty$. For $h \in H$ define spectral functions $F(h; t)$ by

$$F(h; dt) = [1 + h(t)f(t)]f(t). \tag{11.6}$$

Let

(11.7) P_h^n = distribution of the stationary Gaussian sequence $X_1, \ldots, X_n$ when the spectral function is $F(hn^{-1/2}; \cdot)$.

Next, define τ, a linear map of H to $L^2(\mu)$, by

$$(\tau h)(t) = \int_0^t hf^2 \tag{11.8}$$

so that

$$F(h) = F(\theta_0) + \tau h. \tag{11.9}$$

Let B be the closure of $\tau(H)$ in $L^2(\mu)$. Then (τ, H, B) is an abstract Wiener space. Let $\{Q_h, h \in H\}$ be its standard Gaussian shift experiment: we may characterize Q_0 as follows. Let W be standard Brownian motion on the line. Set

$$M(t) = 2\pi \int_0^t f^2(s)\, ds, \quad W_F(t) = W(M(t)), \qquad 0 \le t \le \pi. \tag{11.10}$$

Then Q_0 is the distribution of W_F, a fact easy to check.

Finally, fix α, $1/2 < \alpha < 1$; let $N_c = \{h \in H\colon |h(s) - h(t)| \le c|s-t|^\alpha,\ |h|_\mu \le c\}$. Then $\bigcup_c N_c$ is a dense subspace of H. Define

$$\theta_{nh} = \theta(F(n^{-1/2}h)). \tag{11.11}$$

(11.12) THEOREM. *Under the hypotheses above,*

$$\begin{aligned}\lim_c \liminf_n \inf_{\hat{\theta}_n} \sup_{h \in N_c} \int g(n^{1/2}|\tilde{\theta}_n - \theta_{nh}|)\, dP_h^n &= Eg(|T^{-1} \circ \pi \circ W_F|)\\ &= \lim_c \lim_n \sup_{h \in N_c} \int g(n^{1/2}|\hat{\theta}_n - \theta_{nh}|)\, dP_h^n,\end{aligned}$$

where, as usual, the inf in the first expression is over all estimates of $\{\theta_{nh}\}$ *and* g *is an (bounded) increasing function on the line;* π *as usual is the projection to span* η_i, *and* T *is defined by* (2.10).

(11.13) REMARK. As in the other situations of this paper, the statistical goal here is one of "best fit". A simple model asserts that $X_1, \ldots, X_n$ is stationary Gaussian with spectral function $F(\theta)$. However, it is presumed that the model does not actually fit the data precisely—the data is stationary Gaussian (this hypothesis *can* be weakened, but is outside the scope of this paper) but the actual spectral function is alleged to differ some from $\{F_\theta,\ \theta \in \Theta\}$. Because the model $\{F_\theta\}$ has appealing simplicity, or for other reasons, one still desires to fit it to the data as best one can. Here this goal is formulated technically by (11.2): one wants the spectral function of the form $\{F_\theta,\ \theta \in \Theta\}$ that comes 'closest' to the actual spectral function. It is in this context that the optimality result is to be understood.

PROOF. We begin by showing that the first expression in (11.12) exceeds the second. Define

$$\xi_n(\theta, P_h^n) = F(n^{-1/2}h) - F(\theta).$$

Because of (11.9), this process satisfies (5.19) with V the identity. The hypotheses of identifiability and differentiability in §3 (with derivative η) follow from (11.5); the hypothesis of boundedness is trivial. Moreover, results of Davies (1973) show, via Lemma (4.1), that the experiments $\{P_h^n,\ h \in H_0\}$ converge to $\{P_h,\ h \in H_0\}$, where $H_0 = \bigcup_{c>0} N_c$. The desired inequality is now immediate from (5.20).

Next, let h_n be a sequence in N_c; to finish the proof of the theorem it is necessary (cf. (5.18)ff) only to show that, under $P_{h_n}^n$,

$$n^{1/2}(\hat{\theta}_n - \theta_{nh_n}) \Rightarrow T^{-1} \circ \pi \circ W_F. \tag{11.14}$$

Let $\xi_{1n}(\theta) = \hat{F}_n - F(\theta)$. Arguments of Ibragimov (1963) show that, under $P_{h_n}^n$,

$$(\hat{F}_n - F(n^{-1/2}h_n))n^{1/2} \Rightarrow W_F \quad \text{in } L^2(\mu). \tag{11.15}$$

This easily implies the hypotheses of boundedness for $\xi_{1n}(\theta)$. Since the hypotheses of differentiability and identifiability follow as in the first part of the proof, we have

$$(\hat{\theta}_n - \theta_0) = -T \circ \pi \circ \xi_{1n}(\theta_0) + o(n^{-1/2}) \tag{11.16}$$

by the main result of §5. Similarly, define $\xi_{2n}(\theta) = F(n^{-1/2}h_n) - F(\theta)$; this satisfies the three basic hypotheses too, so

$$(\theta_{nh_n} - \theta_0) = -T \circ \pi \circ \xi_{2n}(\theta_0) + o(n^{-1/2}). \tag{11.17}$$

Then (11.14) is immediate from (11.15)–(11.17). Q.E.D.

XII. Hellinger metric. Let $\{P_\theta,\ \theta \in \Theta\}$ be a family of probabilities, on some Euclidean space S, indexed by Θ, an open subset of R^d. Assume there is a sigma-finite measure μ such that each P_θ is absolutely continuous with respect to μ with density $f_\theta = f(\theta)$:

$$f(\theta;\ ) = dP_\theta/d\mu. \tag{12.1}$$

Let Θ_0 be a subset of Θ of the form

$$\Theta_0 = \Theta \cap M, \tag{12.2}$$

where M is a subspace of R^d having dimension $s < d$. This section treats the problem of trying to fit the model $\{P_\theta,\ \theta \in \Theta\}$ as best as one can using the "simpler" model $\{P_\theta,\ \theta \in \Theta_0\}$. Such a problem arises in certain areas of applied statistics (e.g., multivariate analysis) where the data is probably governed by a parametric model $\{P_\theta,\ \theta \in \Theta\}$ where the dimension of Θ is so large as to be unhelpful in understanding the basic shape of the data; one therefore desires to use a rougher model—with parameter set Θ_0 having much lower dimension—as a workable approximation.

The metric to be used for this study is the Hellinger metric, which is the basic one for much of classical parametric estimation (cf. Le Cam (1973)). The *Hellinger* distance between P_θ and $P_{\theta'}$ is the $L^2(\mu)$ distance between $f_\theta^{1/2}$, $f_{\theta'}^{1/2}$. The relevant minimum distance functional is then $\varsigma(\theta')$ defined by

$$\inf_{\theta \in \Theta_0} |f^{1/2}(\theta') - f^{1/2}(\theta)|_\mu = |f^{1/2}(\theta') - f^{1/2}(\varsigma(\theta'))|_\mu, \tag{12.3}$$

i.e., we naturally want to find $\varsigma(\theta')$, the point in Θ_0 so that $P(\varsigma(\theta'))$ is closest to $P(\theta')$ in the Hellinger metric.

We shall treat the estimation of $\varsigma(\theta)$ under the following hypotheses. Assume for each point $\theta_1 \in \Theta$:

(12.4) (a) If $f^{1/2}(\theta_n)$ converges to $f^{1/2}(\theta_1)$ in $L^2(\mu)$, then $\theta_n \to \theta_1$.

(b) There exists $\eta(\theta_1) = (\eta_1(\theta_1), \ldots, \eta_d(\theta_1))$, $\eta_i(\theta_1) \in L^2\mu$, linearly independent, such that

$$f^{1/2}(\theta) - f^{1/2}(\theta_1) = \langle \theta - \theta_1, \eta(\theta_1)\rangle + o(|\theta - \theta_1|).$$

Fix now $\theta_0 \in \Theta$.

(12.5) Let Γ = matrix with entries $\langle \eta_i, \eta_j\rangle_\mu$, $\eta = \eta(\theta_0) = (\eta_1, \ldots, \eta_d)$. Because η_i are linearly independent, Γ is nonsingular. For convenience we assume

$$(12.6) \qquad \Theta_0 = \Theta \cap M \quad \text{with } M = \{y \in R^d\colon y = (y_1, \ldots, y_s, o, o, \ldots, o)\}.$$

This assumption can always be satisfied by means of an appropriate reparametrization (which can be chosen independently of θ_0).

Let H be the Hilbert space consisting of R^d with the usual Euclidean metric. Let Σ be the matrix

$$(12.7) \qquad \Sigma = \tfrac{1}{4}\Gamma^{-1}.$$

Let B be R^d with norm

$$(12.8) \qquad |x|_B = |\Sigma^{1/2}x|_0,$$

where $|\ |_0$ is Euclidean norm. Define $\tau\colon H \to B$ by $\tau h = h$. Then (τ, H, B) is an abstract Wiener space; if $\{Q_h,\ h \in H\}$ is its Gaussian shift experiment then

$$(12.9) \qquad Q_0 \text{ is } N(0, \Sigma) \text{ on } R^d.$$

For $h \in H$ define

$$(12.10) \qquad P_h^n(dx) = \prod_1^n f(\theta_0 + n^{-1/2}h, x_i)\, dx_i.$$

Define matrices

$$(12.11) \qquad \begin{aligned} \Gamma_s &= \{\langle \eta_i, \eta_j\rangle\colon 1 \le i \le s,\ 1 \le j \le s\}, \\ \Gamma_{sd} &= \{\langle \eta_i, \eta_j\rangle\colon 1 \le i \le s,\ 1 \le j \le d\} \end{aligned}$$

so Γ_{sd} maps R^d to R^s. Then, for any point $x \in R^d$, one may calculate

$$(12.12) \qquad Ax \equiv \Gamma_s^{-1} \circ (\Gamma_{sd}x).$$

Define $N_c = \{h \in H\colon |h| \le c\}$, and let $\hat\theta_n$ be the classical 1-*step maximum likelihood estimator*; this well-known estimator is a bit complicated to describe—see Le Cam (1969).

(12.13) THEOREM.

$$\begin{aligned} \lim_{c\uparrow\infty} \liminf_n \inf_{\tilde\theta_n} \sup_{h\in N_c} \int g(n^{1/2}|\tilde\theta_n - \theta_{nh}|)\, dP_h^n &= \int g(|Ax|)\, Q_0(dx) \\ &= \lim_{c\uparrow\infty} \lim_n \sup_{h\in N_c} \int g(n^{1/2}|\varsigma(\hat\theta_n) - \theta_{nh}|)\, dP_h^n, \end{aligned}$$

where $\theta_{nh} = \varsigma(\theta_0 + hn^{-1/2})$, *and* ς *was defined in* (12.3).

Evidently, the result asserts that $\varsigma(\hat\theta_n)$ *is LAM*.

PROOF. We begin by showing that the first expression in the theorem is at least as big as the second. For this, note first that the experiments $\{P_h^n,\ h \in H\}$ converge to $\{Q_0,\ h \in H\}$. Define

$$\xi_n(\theta, P_h^n) = f^{1/2}(n^{-1/2}h) - f^{1/2}(\theta), \qquad \theta \in \Theta_0.$$

Because of (12.4)(b),

$$n^{1/2}\xi_n(\theta_0, P_h^n) = \langle h, \eta \rangle + o(1) = V \circ \tau h,$$

where $\tau h = h$, $Vx = \langle x, \eta \rangle$. Therefore ξ_n satisfies (5.19). It is also clear from (12.4) that ξ_n satisfies identifiability, boundedness, and differentiability with derivative $\eta_0 = (\eta_1, \ldots, \eta_s, o, o, \ldots, o)$ (this derivative is taken in Θ_0, not Θ). Let $\pi =$ projection in $L^2(\mu)$ to span $\eta_1, \ldots, \eta_s$, and let T be as in (2.10). Then the basic result of §5 asserts that the first expression in (12.13) is bounded below by $\int g(|T^{-1} \circ \pi \circ Vx|)\, P_0(dx)$. Using the evaluation of $T^{-1} \circ \pi$ given in (2.19), and the explicit form of V above, one obtains from this the expression given in the theorem.

To finish the proof it is now necessary to show only that if h_n is a sequence in N_c, then $n^{1/2}[\varsigma(\hat{\theta}_n) - \varsigma(\theta_{nh_n})]$ converges under $P_{h_n}^n$, to $T^{-1} \circ \pi \circ V \circ Z$, where Z has distribution P_0 (see (5.18)ff). To do this, use the well-known fact that

$$n^{1/2}[\hat{\theta}_n - (\theta_0 + h_n n^{-1/2})] \Rightarrow Z; \tag{12.14}$$

see Le Cam (1969) or Roussas (1972).

Next, set $\xi_{1n}(\theta) = f^{1/2}(\hat{\theta}_n) - f^{1/2}(\theta)$; this satisfies, under $P_{h_n}^n$, the hypotheses of differentiability (with derivative η_0), identifiability, and boundedness (because of (12.14) and the fact that h_n is bounded). Therefore, using (12.4)(b) and (12.14),

$$\begin{aligned} n^{1/2}[\varsigma(\hat{\theta}_n) - \theta_0] &= -T^{-1} \circ \pi \circ \xi_{1n}(\theta_0) + o(1) \\ &= -T^{-1} \circ \pi \circ V(\hat{\theta}_n - \theta_0) + o(1). \end{aligned} \tag{12.15}$$

On the other hand, if $\xi_{2n}(\theta) = f^{1/2}(\theta_0 + h_n n^{-1/2}) - f^{1/2}(\theta)$ the three hypotheses are again satisfied, so

$$n^{1/2}(\theta_{nh_n} - \theta_0) = -T^{-1} \circ \pi \circ V(n^{-1/2}hn). \tag{12.16}$$

The desired result is now immediate from (12.14)–(12.16).

(12.17) REMARK. It is of interest to consider a somewhat different estimator: to estimate $\theta_{nh} \in \Theta_0$, first calculate $\hat{\theta}_n$, the one step MLE for the big model $\{P_\theta,\ \theta \in \Theta\}$; then estimate θ_{nh} by the first s coordinates of $\hat{\theta}_n$. If $\eta_{s+1}, \ldots, \eta_d$ are orthogonal to $\eta_1, \ldots, \eta_s$, then facts given in the foregoing proof easily show that such an estimator is also LAM. On the other hand, if the aforementioned orthogonality fails, the newly proposed estimator and $\varsigma(\hat{\theta}_n)$ are not asymptotically equivalent (and the new one will not be LAM in the present framework).

It is possible, by a reparametrization, to make $\eta_{s+1}, \ldots, \eta_d$ orthogonal to $\eta_1, \ldots,$ η_s. This reparametrization will change the definition of what is being estimated, and, moreover, the *reparametrization will depend on* θ_0. Evidently, the statistical meaning of a local asymptotic investigation is completely destroyed if the definition of what is being estimated is changed for the convenience of the point θ_0 chosen for the local analysis.

XIII. Further methods in the independent case. Let Θ be an open subset of R^d and $\{P_\theta,\ \theta \in \Theta\}$ a family of probabilities on R^s. Let $X_1, \ldots, X_n$ be independent random R^s-valued random variables with common distribution G. As in other sections, the goal is to 'fit' the model $\{P_\theta\}$ as best one can to the actual data distribution. In §6 this was done by matching up cdf's as close as one could; in §9 we "fitted" the model to the data by matching the quantiles. Many other methods are possible and have practical importance. This section exhibits some of the possibilities together with natural LAM estimators.

We begin by giving the relevant abstract structure; many illustrations of it follow.

Let $(Y, \mathcal{Y})$ be a measure space, μ a measure on $\mathcal{Y}$. Y will be assumed Euclidean, but this is not really necessary. Let ς be a mapping defined on probabilities on R^s such that if G is a probability, then $\varsigma(G)$ is a real measurable function on Y: its value at $t \in Y$ is $\varsigma(G)(t)$. ς need not be defined on all probabilities—only on the probabilities P_θ and a certain few others to be described below. We assume that $\varsigma(G)$ satisfies

$$\int [\varsigma(G)(t)]^2\, \mu(dt) < \infty, \tag{13.1}$$

so ς is an $L^2(\mu)$-valued functional defined on certain measures. Define the functional $\theta(G)$ by

$$\inf_\theta |\varsigma(G) - \varsigma(P_\theta)|_\mu = |\varsigma(G) - \varsigma(P_{\theta(G)})|_\mu, \tag{13.2}$$

where, of course, it is assumed that ς is defined on G. The statistical problem is to estimate $\theta(G)$.

The basic hypothesis on $\varsigma(G)$ is that it is *essentially* given *locally* by a recipe of the type $\varsigma(G)(t) = \int \psi(t,x)\, G(dx)$ for some measurable function ψ on $Y \times R^s$. This will be described next, but it takes a bit of time. The classic M, L and R functionals have this simple structure, but the class to be described is much broader, including even the functionals of §§6, 9.

The probabilities P_θ, which are defined on the Borel sets of R^s, are assumed absolutely continuous with respect to a sigma-finite (nonatomic) measure ν. As in previous LAM developments, fix θ_0. Let

$$f(\theta_0, x) = \text{density of } P_{\theta_0}. \tag{13.3}$$

Let

$$\begin{aligned} &\psi(t,x) \text{be a real (or complex) jointly} \\ &\text{measurable function on } Y \times R^s. \end{aligned} \tag{13.4}$$

This function depends on θ_0 also. Assume that, for each $t \in y$,

(13.5) (a) $\int \psi^2(t,x)\, dP_{\theta_0}(x) < \infty$,

(b) $\int \psi(t,x) f(\theta_0, x)\, \nu(dx) = 0$.

Define H to be the closure (in $L^2(P_{\theta_0})$) of span $\{\psi(t,\cdot)\colon t \in Y\}$. Note that any ψ that satisfies (13.5)(a) can be made to satisfy (13.5)(b) merely by subtracting from it the function of t given in (b). Note also that H here is a sub-Hilbert space of the one given in §6. Define $\tau\colon H \to L^2(Y, \mu)$ by

$$(\tau h)(t) = \langle \psi(t, \cdot), h\rangle_H. \tag{13.6}$$

Note that τ is one-to-one on H, since if $\langle\psi(t;\cdot),h\rangle_H = 0$ for all t, then $h \in H$ is orthogonal to H. Define

(13.7) $$B = \text{closure of } \tau H \text{ in } L^2(\mu).$$

Define the mapping $K(s,t)$, $s,t \in Y$, by

(13.8) $$K(s,t) = \langle\psi(t;\cdot),\psi(s;\cdot)\rangle_H.$$

Then, since τ is a Hilbert-Schimdt operator, (τ, H, B) is an abstract Wiener space; if $\{Q_h,\ h \in H\}$ is its Gaussian shift experiment, then Q_0 is a Gaussian measure on $L^2(Y,\mu)$ with covariance operator S given by

(13.9) $$(Sg)(t) = \int K(s,t)g(s)\,\mu(ds).$$

Alternately, we can characterize Q_0 as follows. Let $X_1,\dots,X_n$ be i.i.d. random variables with distribution P_{θ_0}; then $\{\psi(\cdot;X_i)\}$ is an i.i.d. sequence of $L^2(\mu)$-valued random variables. A routine application of the central limit theorem for Hilbert space (cf., Parthasarathy (1967)), shows that

(13.10) $n^{-1/2}\Sigma\psi(\cdot;X_i)$ converges in distribution (on $L^2(\mu)$) to a random variable having distribution Q_0.

For $h \in H$ define probabilities P_h by specifying that the density of P_h is $f(h;\cdot)$ given by

(13.11) $$f(h;x) = f(\theta_0;x)[1 + h(x)].$$

Let

(13.12) $$P_h^n = n\text{-fold product measure of } P_{n^{-1/2}h}.$$

Then $\{P_h^n\}$ converges to $\{Q_h\}$. Let H_0 be a dense subspace of H.

We can now introduce our hypotheses on the functional ς and the family $\{P_\theta\}$. Assume first, for each $h \in H_0$,

(13.13) $$\begin{aligned}&\lim_{n\to\infty}[\varsigma(P_{n^{-1/2}h})(t) - \varsigma(P_{\theta_0})(t)]n^{1/2}\\ &\quad= \int \psi(t;x)h(x)f_{\theta_0}(x)\,\nu(dx) \equiv (\tau h)(t),\end{aligned}$$

convergence in $L^2(Y,d\mu)$.

This hypothesis is *much* weaker than the hypothesis that the functional $G \to \varsigma(G)$ be Gateaux differentiable at G. In most applications ς is ***not even defined*** on all G; we need only that ς be defined on $\{P_\theta\}$ and on the measures $P_{n^{-1/2}h}$, $h \in H_0$—a considerably smaller class.

Assume further: there exists $\eta = (\eta_1,\dots,\eta_d)$, $\eta_i \in L^2(Y,\mu)$, such that

(13.14) $$\varsigma(P_\theta) - \varsigma(P_{\theta_0}) = \langle\theta - \theta_0, \eta\rangle + o(|\theta - \theta|).$$

Often there is a vector $h_0 = (h_1,\dots,h_d)$ such that, for θ close to θ_0, $f_\theta \doteq f_{\theta_0}[1 + \langle\theta - \theta_0, h_0\rangle]$; this is true, for example, when $\{P_\theta\}$ is quadratic mean differentiable. When such an approximation holds, a variant of (13.13) often holds as well (even though $\langle\theta - \theta_0, h_0\rangle$ will usually *not* belong to H_0), and in this case $\langle\theta - \theta_0, \eta(t)\rangle = \int\psi(t;x)\langle\theta - \theta_0, h_0\rangle f_{\theta_0}(x)\,\nu(dx)$, an explicit form for η. Finally, assume:

(13.15) If $\varsigma(P_{\theta_n})$ converges to $\varsigma(P_{\theta_0})$ in $L^2(Y,\mu)$, then $\theta_n \to \theta_0$.

Let N_c, $c > 0$, be defined by $N_c = \{h \in H_0\colon |h| \le c\}$. Let π be the projection in $L^2(\mu)$ to span η_i, and define T as in (2.10).

(13.16) THEOREM (LAM LOWER BOUND). *For any increasing function g,*

$$\lim_{c}\liminf_{n}\inf_{\tilde{\theta}_n}\sup_{h\in N_c}\int g(n^{1/2}|\tilde{\theta}_n - \theta_{nh}|)\,dP_h^n \geq \int g(|T^{-1}\circ \pi x|)\,Q_0(dx),$$

where $\theta_{nh} = \theta(P_{n^{-1/2}h})$.

(13.17) REMARK. The neighborhoods $\{P_h^n : h \in N_c\}$ can be replaced by a *bigger* collection of product measures $\{G^n\}$ and the theorem remains true (of course dP_h^n is replaced by dG^n, θ_{nh} by $\theta(G)$, and so forth).

PROOF. If one defines $\xi(P_h^n, \theta) = \varsigma(P_{n^{-1/2}h}) - \varsigma(P_0)$, then

$$n^{1/2}\xi_n(P_h^n, \theta_0) = n^{1/2}\xi_n(P_0^n, \theta_0) + \tau h + o(1),$$

so (5.19) holds. The hypotheses of boundedness, identifiability, and differentiability hold because of (13.13)–(13.15), so the result is immediate from the development of §5.

Let $\hat{F}_n(dx)$ be the empirical measure of $X_1, \ldots, X_n$:

(13.18) $$\hat{F}_n(A) = n^{-1}\Sigma I_A(X_i), \qquad A \text{ a Borel set of } R^s.$$

Make the assumption:

(13.19) If h_n is a sequence in N_c, then

$$n^{1/2}[\varsigma(\hat{F}_n) - \varsigma(P_{n^{-1/2}h_n})] \Rightarrow Q_0 \quad \text{under } P_{h_n}^n,$$

$$n^{1/2}[\varsigma(P_{n^{1/2}h_n}) - \varsigma(P_{\theta_0})] = \tau h_n + o(1).$$

Except for regularity problems, one would expect such a hypothesis to hold. Indeed, one needs 'only' that $\varsigma(\hat{F}_n) - \varsigma(P_{n^{-1/2}h_n}) \doteq n^{-1/2}\Sigma\psi(\cdot, X_i)$, which is a variant of (13.13) (i.e., one needs the 'Gateaux derivative' in 'direction' $\hat{F}_n$ and in direction $P_{n^{-1/2}h_n}$), and then (13.19) follows from (13.10). Under these additional assumptions, $\theta(\hat{F}_n)$ is LAM:

(13.20) THEOREM. *If g is bounded and uniformly continuous, then*

$$\lim_{c}\lim_{n}\sup_{h\in N_c}\int g(n^{1/2}|\theta(\hat{F}_n) - \theta_{nh}|)\,dP_h^n = \int g(T\circ \pi x)\,dQ_0.$$

(13.21) REMARK. The proof, as usual, proceeds by showing that if $h_n \in N_c$, then $n^{1/2}(\theta(\hat{F}_n) - \theta_{nh_n})$ converges to $T\circ\pi\circ Z$, $Z \sim P_0$: i.e., a convergence to the proper limit that is *uniform* over N_c. In many applications the neighborhoods N_c are very skimpy. In such cases it is important to prove that this convergence is uniform over somewhat broader neighborhoods (not necessarily parametrized only by H). This will not change the LAM lower bound (cf. (13.17)) and will enhance the desirability of the estimator $\theta(\hat{F}_n)$. The fact that *some* skimpy neighborhoods *cannot* be extended much brings up serious questions concerning the ability of the selected minimum distance method to deal with data that suffers from moderately severe departures from the model.

We now discuss a number of examples. We dwell mainly on (13.13) and on the form of ς, ψ, H, τ, so that the general structure becomes apparent; in most cases the identifiability and differentiability hypotheses will be assumed.

EXAMPLE 1. Let A_t be the set in R^s given by $A_t = \{u \in R^s : u_i \leq t_i\}$ when $t = (t_1, \ldots, t_s) \in R^s$. Define for any probability G on R^s, $\varsigma(G)(t) = G\{A_t\}$. Let μ

be a finite measure on R^s; then $\varsigma(G) \in L^2(\mu)$. Fix a parametric family of measures $\{P_\theta\}$, and fix θ_0. Then

$$|\varsigma(G) - \varsigma(P_\theta)|_\mu^2 = \int_s [G(A_t) - P_\theta(A_t)]^2 \, d\mu,$$

and so this set up is designed for the situation when 'best fit of model to data' is defined by matching up the 'cumulatives' as closely as possible. Of course, if $s = 1$, §6 treats a special case. To put the present case in the framework of this section, define on $R^s \times R^s$ the mapping

$$\psi(t,x) = I_{A_t}(x) - P_{\theta_0}(A_t).$$

It is easy to see that then

$$H = \left\{ h \in L^2(dP_{\theta_0}) \colon \int h \, dP_{\theta_0} = 0 \right\} = \operatorname{span}\{\psi(t,\cdot) \colon t \in R^s\}.$$

Then

$$(\tau h)(t) = \langle \psi(t,\cdot), h \rangle = \int_{A_t} h(u) f(\theta_0, u) \, \nu(du), \qquad h \in H.$$

So τh is actually a continuous function. Moreover, for *any* G,

$$\varsigma(G)(t) - \varsigma(P_{\theta_0})(t) = \int \psi(t,x) \, d(G - P_{\theta_0})(x),$$

so (13.13) is an *identity* for each n. If $N_c = \{h \colon |h| \leq c\}$, then for this choice both theorems hold. In fact, more is true: if G_n is a sequence of probabilities such that $\int |G_n(A_t) - P_{\theta_0}(A_t)|^2 \, \mu(dt) \leq cn^{-1/2}$, and if $X_1, \ldots, X_n$ are i.i.d. G_n, then $n^{1/2}[\hat{F}_n(A_t) - G_n(A_t)]$ converges in distribution on $L^2(\mu)$, implying that $\theta(\hat{F}_n)$ converges uniformly over these neighborhoods to its limit. That is, one can take greatly extended neighborhoods of P_{θ_0}, as suggested in the remarks above.

EXAMPLE 2. Here is a variant of Example 1: suppose one wants to fit the model to the data by matching up the characteristic functions as closely as possible (instead of matching up the cumulatives). From a practical point of view one probably wants to estimate a few probabilities, and the method just suggested might not be too helpful for that; but it is conceivable that someone would want to estimate a characteristic function. Assuming this, one may place the problem in the general framework of this section as follows: Define, for any probability G on R^s, the map $\varsigma(G)(t) = \int \exp i\langle t,x \rangle \, G(dx)$. Let μ be a finite measure on R^s; then $\varsigma(G) \in L^2(\mu)$ (complex L^2 space). Fix a parametric family $\{P_\theta,\ \theta \in \Theta\}$ and $\theta_0 \in \Theta$. Then $|\varsigma(G) - \varsigma(P_\theta)|$ is the $L^2(\mu)$ distance between the characteristic functions of G, P_θ. Define on $R^s \times R^s$ the map

$$\psi(t,x) = \exp i\langle t,x \rangle - \int \exp i\langle t,x \rangle \, P_{\theta_0}(dx).$$

It is complex valued, and we decline to identify H (it is not hard), the span of $\{\psi(t,\cdot) \colon t \in R^s\}$ in $L^2(dP_{\theta_0})$. The map τ, in the present case, turns out to be

$$(\tau h)(t) = \int \exp\{i\langle t,x \rangle\} h(x) \, dP_{\theta_0}(x), \qquad h \in H.$$

Again, for any G,

$$\varsigma(G)(t) - \varsigma(P_{\theta_0})(t) = \int \psi(t,x)\, d(G - P_{\theta_0})(x)$$

so (13.13) is an identity for each n. For reasons given in the preceding example, $\theta(\hat{F}_n)$ here is again LAM, the uniform convergence holding over broad neighborhoods like those suggested in that example.

EXAMPLE 3. Suppose the data distribution is G (unknown probability on the line), and we want to fit the model $\{P_\theta\}$ to the data by choosing θ such that the *mean* of P_θ is closest to the mean of G. This somewhat degenerate situation is discussed briefly here, in preparation for Example 4 (when extended to regression problems, it becomes, essentially, the well-known least squares method). One could use pth-moment instead of mean—the discussion is essentially the same.

To put the foregoing into our framework, define $\varsigma(G) = \int x\, G(dx)$. Note that ς is not defined for very many G. In this case $L^2(Y, \mu)$ will be the real line with the usual distance. As usual, fix θ_0. Define $\psi(t,x)$ on $R' \times R'$ by $\psi(t,x) = x - c$ for all t, where $c = \int x\, dP_{\theta_0}(x)$. Then H is the one-dimensional subspace of $L^2(dP_{\theta_0})$: $H = \{\lambda(x-c) \colon \lambda \in R'\}$. The mapping τ is $\tau\{\lambda(\cdot - c)\} = \lambda\sigma^2$, where $\sigma^2 = \int (x-c)^2\, P_{\theta_0}(dx)$; here $\lambda(\cdot - c)$ is a generic element of H. The measure P_0 is $N(0, \sigma^2)$. The differentiability assumption becomes the ordinary differentiability at θ_0 of $\theta \to \int x\, dP_\theta(x)$; nonsingularity just means that the value of the derivative at θ_0 is not zero. Identifiability means essentially that different P_θ's must have different means. Again, in this example, (13.13) is trivial, holding without the need of taking any limit.

It is important to notice that here our usual neighborhoods $N_c = \{h \in H \colon |h| \leq c\}$ are *one dimensional.* Assuming $\int (x-c)^2\, P_{\theta_0}(dx) < \infty$, the normalized sample mean converges, under $P^n_{h_n}$, $h_n \in N_c$, to P_0. Since in the present case π is the identity, the minimum distance estimate of θ is $\overline{X}$, the sample mean. The development so far is unsatisfactory, since the choice of N_c dictated by (13.15)ff is only one dimensional, so the allowable data distributions are extremely limited (and most likely do not include any P_θ's other than P_{θ_0}!). Therefore, to get a useful result, one must attempt to show that $\overline{X}$ converges over rather broad neighborhoods of P_{θ_0}. If one attempts to use the N_c of, say, Example 1 then such convergence will fail: typically, measures of the form $f_{\theta_0}(1 + hn^{-1/2})$, where h satisfies only $\int f_\theta h = 0$, will *not even have a mean*; so the "neighborhoods" of Example 1 are too big. An intermediate neighborhood N_c (that works!) is of the form $\{h\colon \int hf_\theta = 0$, support $h \subset [-c, c]$, $|h| \leq c\}$. This will not be completely satisfactory, since it will still probably not contain P_θ for θ close to θ_0. Another possibility is $\{h\colon \int hf_\theta = 0$, $|h|^2 \leq c$, $\int x^2h^2 f_{\theta_0} \leq c\}$.

EXAMPLE 4. This example presents a very reasonable minimum distance method for the situation where the data distribution is believed to be not too different from those posited by the model $\{P_\theta\}$. Suppose the measures are on the line. The idea is to find θ so that the distance between the cdf's and first k moments of P_θ, G is as small as possible: that is, find θ that achieves the inf:

$$\inf_\theta \left\{ |G - P_\theta|^2 + \sum_{i=1}^{k} \left[\int x^i\, dG - \int x^i\, dP_\theta \right]^2 \right\},$$

where $|G - P_\theta|$ is an L^2-norm between the cdf of G, P_θ. We show next how to formulate this using the structure developed earlier in this section. Let $e_0, e_1, \ldots, e_k$ be an orthonormal set in $L^2([0,1], dx)$ (unit interval with Lebesgue measure). Let $G(t), P_\theta(t)$ denote cdf's of G, P_θ. Fix θ_0. Let m be a finite measure on the line, so $G(t) \in L^2(m)$. Define $Y = [0,1] \times R^1$, $d\mu = dx \times dm$. The $L^2(Y; d\mu)$-valued elements ς will now be defined by

$$\varsigma(G)(t_1, t_2) = e_0(t_1)G(t_2) + \sum_{i=1}^{k} e_i \int x^i \, dG(x).$$

Then

$$|\varsigma(G) - \varsigma(P_\theta)|_\mu^2 = \int [G(t) - P_\theta(t)]^2 \, m(dt) + \sum \left[\int x^i \, d(G - P_\theta) \right]^2,$$

so the θ that achieves $\inf_\theta |\varsigma(G) - \varsigma(P_\theta)|$ is the one we want. Define, for $t = (t_1, t_2) \in Y$,

$$\psi(t, x) = e(t_1)[I_{t_2}(x) - P_{\theta_0}(t_2)] + \sum_1^k e_i(t_1)(x^i - c_i),$$

where $c_i = \int x^i P_{\theta_0}(dx)$. It is clear that $H = \{h\colon \int h^2 \, dP_{\theta_0} < \infty, \int h \, dP_{\theta_0} = 0\}$, and it is fairly routine to find τ, P_0, etc., for this particular example; as in all other cases so far (13.13) is trivial. For reasons given in Example 3, one cannot use $N_c = \{h \in H\colon |h| \le c\}$ here: $\varsigma(P_{hn^{-1/2}})$ is not defined for most of these h. One workable possibility is $N_c = \{h \in H\colon |h| \le c, \text{ support } h \subset [-c, c]\}$. Then N_c increases with c, N_c is convex, and $\bigcup N_c$ is dense in H. With the foregoing choice (assuming as usual the differentiability and identifiability hypotheses), $\theta(\hat{F}_n)$ is LAM.

EXAMPLE 5. In this example let us try to fit the model to the data by matching up quantiles. This was done in detail in §9; here, as preparation for Example 6, we indicate only how the result of §9 fits into the present framework. We do not dwell on regularity hypotheses; for these, see §9.

Define, for a distribution G on the line, $\varsigma(G)(t) = G^{-1}(t) I_{[\alpha,\beta]}(t)$, where $t \in [0,1]$, $0 < \alpha < \beta < 1$. Let $Y = [\alpha, \beta]$ and let μ be a finite measure on Y; then $\varsigma(G) \in L^2(Y, d\mu)$ (assuming support G is sufficiently large—see §9). Bring in the parametric family $\{P_\theta\}$ of probabilities on the line and fix θ_0. Define on $[\alpha, \beta] \times R$ the map

$$\psi(t, x) = c_t [I_{A_t} - t],$$

where

$$c_t = 1/f(F^{-1}(t)), \quad A_t = (-\infty, F^{-1}(t)], \quad F = \text{cdf of } P_{\theta_0}.$$

It is clear that

$$H = \operatorname{span}\{\psi(t, \cdot),\ t \in [\alpha, \beta]\} = \left\{h\colon \int h^2 f(\theta_0) < \infty, \int h f_{\theta_0} = 0\right\}.$$

The map τ is given by

$$(\tau h)(t) = c_t \int^{F^{-1}(t)} h(u) f_{\theta_0}(u) \, du.$$

The covariance kernel

$$K(s,t) = \langle \psi(t,\cdot), \psi(s,\cdot)\rangle = c_t c_s[s \wedge t - st]$$

on $[\alpha,\beta] \times [\alpha,\beta]$ and this identifies P_0 as the distribution of $c_t W^0(t)$, $\alpha \le t \le \beta$, where W^0 is the usual Brownian bridge on $[0,1]$. Finally, note that if $h \in H$ is sufficiently smooth,

$$\lim_n n^{1/2}[\varsigma(P_{h^{-1/2}h}) - \varsigma(P_{\theta_0})] = \tau h; \tag{13.22}$$

see §. This is the first example where (13.13) is not an identity. As pointed out in §9, the neighborhood system N_c has to be chosen with some care to ensure (13.13) for $h \in N_c$. That $\theta(\hat{F}_n)$ is here optimal was discussed in §9.

EXAMPLE 6. In this example we merely point out a number of interesting possibilities based on examples above.

(a) Instead of basing the notion of 'best fit' on the quantile *function* as in Example 6, one could pick a particular quantile (e.g., the median) and define θ to be estimated as the one that makes (say) median P_θ closest to median G, where G is the unknown data distribution. Evidently this is the analogue for quantile functions of Example 3. More generally, one may proceed as follows. Let Y be some space, and for each $t \in Y$, let $a(t,dx)$ be a measure on the line. The proposal then is to base a minimum distance method on the elements $\varsigma(G)(t) = \int G^{-1}(u)\,a(t;du)$, where G is a cdf on the line. If a is independent of t, then ς is the usual *L-functional*, common in robustness studies ($a(t,dx)$ = unit mass at 1/2 gives median, for example). Many interesting possibilities arise on letting $a(t,dx)$ vary with t ($a(t,dx)$ = unit mass at $\{t\}$ gives Example 6). The methods of this section are easily adapted to analyze such functionals.

(b) Picking up one of the suggestions of Example 6(a), we can return to Examples 1–3 and introduce *kernels* $k(t,x)$, where t is in some appropriate space, and define for a probability distribution G, $\varsigma(G)(t) = \int k(t,x)\,G(dx)$. The choices $k(t,x) = I_{A_t}(x)$, $A_t = \{u\colon u \le t\}$, $k(t,x) = \exp\{i\langle t,x\rangle\}$, and $k(t,x) = x^P$ yield, respectively, Examples 1–3. More fun is possible on letting t vary in appropriate infinite-dimensional Hilbert spaces. Of course, $\varsigma(G)(\cdot)$ should belong to some $L^2(\mu)$-space. As another variant, one may define ς to be a favorite M-functional defined on distributions G (not necessarily distributions on the line). The minimum distance methods based on these choices of ς are easily analyzed by the theory of this section.

(c) Finally, one may desire to fit the model to the data by matching up *both* the cdf's and the quantiles. Here is one possibility. Let $t = (t_1,t_2,t_3)$ denote a point in the space $Y = R' \times [\alpha,\beta] \times [0,1]$, where $0 < \alpha < \beta < 1$. Let m_1 be a measure on R', m_2 a measure on $[\alpha,\beta]$ and set $\mu = m_1 \times m_2 \times m_3$, m_3 = Lebesgue measure on $[0,1]$. Define, for any probability F,

$$\varsigma(F)(t) = F(t_1)e_1(t_3) + F^{-1}(t_2)e_2(t_3),$$

where e_1, e_2 are orthonormal in $L^2([0,1])$. Then if $|\ |_\mu$ denotes $L^2(\mu)$-norm,

$$|\varsigma(F)|_\mu^2 = \int F^2(t_1)\,dm_1(t_1) + \int [F^{-1}(t_2)]^2\,m_2(dt_2).$$

The theory of this section then easily applies to the minimum distance procedure based on $\varsigma(F) \in L^2(\mu)$.

EXAMPLE 7. Up to now we have assumed Y is a Euclidean space. The theory can be pressed further; for illustration, let us take Y to be $L^2(R', dm)$. A modest amount of care is needed to carry through the forthcoming suggestion rigorously, but let us proceed heuristically. Let us define a functional ς on (certain) probabilities G on the line by $\varsigma(G)(t) = \int t(u)\, G(du)$ if $t \in Y$. This will, of course, not be defined for all G—it depends on what m is. Let $\{e_i\}$ be an orthonormal set in Y and let μ be the Gaussian measure on Y which is the distribution of $\Sigma\lambda_i X_i e_i$, where $\Sigma\lambda_i^2 < \infty$ and the X_i are i.i.d. $N(0,1)$. Readers familiar with the theory of Gaussian measures on Hilbert spaces can give more elegant descriptions of μ. In the formalism of this section, the proposed $\psi(t,x) = t(x)$; there are problems here with joint measurability, but never mind. If $\{P_\theta\}$ is our usual parametric family, then the minimum distance problem is to estimate $\theta(G)$, the point which achieves the infimum in

$$\inf_\theta |\varsigma(G) - \varsigma(P_\theta)|_\mu^2 \doteq \inf_\theta \sum \lambda_1^2 \left[\int e_i \, d(P_\theta - G)\right]^2 .$$

It is probably easier to analyze this metric directly, rather than representing it in ψ-form; nevertheless, we can easily believe now that asymptotic normality and LAM results do indeed hold in the present case. The details are not difficult, given the structural results of §§3, 5.

XIV. Estimating a mixture. Except for technicalities, the example of this section is conceptionally much simpler than that of §6. It is included to illustrate the usefulness of a parameter set Θ which is an infinite-dimensional Hilbert space.

To describe the problem let $F(t,u)$ be a measurable function of $R^1 \times R^d$ such that for each u, $t \to F(t,u)$ is a cdf of the line. The function F is assumed known. If ν is a probability on R^d, define the "mixture" $F(t,\nu)$ by

$$F(t,\nu) = \int F(t,u)\, \nu(du), \tag{14.1}$$

so $F(\cdot,\nu)$ is again a cdf on R^1. Let $X_1, \ldots, X_n$ be i.i.d. with common distribution given by $F(\cdot,\nu)$ for some ν. The statistical problem is to estimate the mixing distribution ν. This section gives a LAM solution to a subset of this problem using minimum distance methods. Since the technicalities are lengthy, we illustrate only the general approach.

To describe the approach first bring in a known sigma-finite measure μ on R^d. For a given mixing distribution ν, let θ be its cdf on R^d:

$$\theta(a) = \nu\{u \colon u \leq a\}. \tag{14.2}$$

Identify ν with θ and assume that the exact set of available mixed cdf's is given by the recipe

$$F(t,\theta) = \int (\theta)(a)\psi(t,a)\, \mu(da) + g(t), \tag{14.3}$$

where ψ is a known jointly measurable function on $R^1 \times R^d$, g is a known function on R^1, and θ is some point in $L^2(R^d,\mu)$. In many examples, (14.3) is obtained from (14.1) by an integration by parts, and μ becomes either Lebesgue measure or a counting measure; nevertheless, (14.3) typically restricts both $F(t,u)$ and the available mixing measures. For convenience, we assume $g = 0$. Also assume that $T\colon \theta \to F(\cdot, \theta)$ is a continuous linear operator into $L^2(m)$, where m is some sigma-finite measure on R^1. With these assumptions, the proposed estimate $\hat{\theta}_n$ is the minimum distance estimator determined by

$$\inf_{\theta \in L^2(\mu)} |\hat{F}_n - T(\theta)|_m = |\hat{F}_n - T(\hat{\theta}_n)|_m, \tag{14.4}$$

where $|\ |_m$ is the norm of $L^2(m)$, and $\hat{F}_n$ is the empirical cdf. This estimate has been studied by Burman (1983), who used methods from integral equations; we develop instead our slightly more general approach using the structural theorems of §§3, 5.

Because of (2.13) and the linearity of T, it is immediate that

$$\hat{\theta}_n = K^{-1} T^* \hat{F}_n, \tag{14.5}$$

provided we assume K^{-1} exists. In the present situation,

(14.6)
$$\begin{aligned}
(T\theta)(t) &= \int \theta(u)\psi(t,u)\,\mu(du),\\
(T^*V)(a) &= \int V(t)\psi(t,a)\,m(dt),\\
(K\theta)(s) &= \int \theta(a)U(a,s)\,\mu(da), \quad \text{where } U(a,s) = \int \psi(t,a)\psi(t,s)\,m(dt).
\end{aligned}$$

If θ_0 is fixed and $X_1, \ldots, X_n$ are i.i.d. $T\theta_0$, then (14.5) implies that

$$n^{1/2}(\hat{\theta}_n - \theta_0) \Rightarrow Y \tag{14.7}$$

in distribution on $L^2(\mu)$. Here Y is a Gaussian random element of $L^2(\mu)$ with covariance operator specified by (2.15), with S there given by the covariance operator of $n^{1/2}(\hat{F}_n - F(\cdot, \theta_0))$. Thus the minimum distance estimate is asymptotically normal.

To complete this section we sketch briefly the LAM framework. Fix cdf θ_0 having density $f\colon \theta_0(t) = \int^t f$. Let $H = \{h\colon \int fh = 0\}$, $\tau_0 h(t) = \int^t hf$, $\tau h(t) = \int \psi(t,u)(\tau_0 h)(u)\,\mu(du)$, $\tau\colon h \to L^2(m)$. Simple considerations show that $(\tau, H, \overline{\tau H})$ is an abstract Wiener space, P_0 being the distribution of the process TY of (14.7). If the local perturbations of the distributions of the X_i at time n are

$$\int \psi(t,u)(\theta_0 + n^{-1/2}\tau_0)(u)\,\mu(du),$$

then one obtains the LAM property as in §6.

XV. Proof of (2.8), (3.6).

PROOF OF (2.8). The first point in the proof is that

$$\inf_{|\theta| \le c} |\xi_n(\theta)| = \inf_{\theta} |\xi_n(\theta)| \tag{15.1}$$

for all n, with probability approaching 1 as $c \to \infty$. To see this, note that since $|\xi_n(\theta)| \ge |\xi_n(\theta) - \xi_n(\theta_0)| - |\xi_n(\theta_0)|$, it follows that

$$\inf_{|\theta| > c} |\xi_n(\theta)| \ge \inf_{|\theta| > c} |\xi_n(\theta) - \xi_n(\theta_0)| - |\xi_n(\theta_0)|.$$

Because of (2.1), (2.4), the right side of this expression goes to $+\infty$ in probability as $c \uparrow \infty$, proving (14.1).

Next, fix c (large) and let $\hat{\theta}_n$ satisfy

$$\inf_{|\theta|\le c} |\xi_n(\theta)| = |\xi_n(\hat{\theta}_n)|. \tag{15.2}$$

Since $|\hat{\theta}_n - \theta_0| < 2c$,

$$\xi_n(\hat{\theta}_n) = \xi_n(\theta_0) + \delta_n\langle\hat{\theta}_n - \theta_0, \eta\rangle + o(1) \tag{15.3}$$

by (2.2). It is easy to see that, for all large c,

$$\inf_{|\theta-\theta_0|<c} |\xi_n(\theta_0) + \delta_n\langle\theta - \theta_n, \eta\rangle| = \inf_{\theta} |\xi_n(\theta_0) + \delta_n\langle\theta - \theta_0, \eta\rangle|, \tag{15.4}$$

so any point $\tilde{\theta}_n$ that achieves the inf in (14.4) satisfies

$$\delta_n\langle\tilde{\theta}_n - \theta_0, \eta\rangle = -\pi\xi_n(\theta_0), \qquad |\tilde{\theta} - \theta_0| \le c, \tag{15.5}$$
$$\xi_n(\tilde{\theta}_n) = (1-\pi)\xi_n(\theta_0) + o(1).$$

By (14.1) and differentiability,

$$|\xi_n(\hat{\theta}_n)| = |\xi_n(\tilde{\theta}_n)| + o_n(1). \tag{15.6}$$

By the Pythagorean theorem,

$$|\langle\hat{\theta}_n - \theta_0, \eta\rangle - \langle\tilde{\theta}_n - \theta_0, \eta\rangle|^2 = |\xi_n(\hat{\theta}_n)|^2 - |\xi_n(\tilde{\theta}_n)|^2 + o(1) = o(1). \tag{15.7}$$

This implies

$$\xi_n(\hat{\theta}_n) = \xi_n(\tilde{\theta}_n) = (1-\pi)\xi_n(\theta_0) + o(1) \tag{15.8}$$

and

$$\delta_n\langle\hat{\theta}_n - \theta_0, \eta\rangle = -\pi\xi_n(\theta_0) + o(1),$$

which is the desired result.

PROOF OF (3.6). As remarked in §3, one proves (3.6) by applying (2.8) to the processes $\xi_n^0(\theta) = n^{1/2}[\xi_n(\theta_0 + \theta n^{-1/2})]$. One need only check the differentiability, boundedness, and identifiability hypotheses of §2 when applied to $\xi_n^0(\theta)$. We indicate here only identifiability, since the other two are clear. For this take $\theta_0 = 0$ for simplicity and suppose that $\inf_n \inf_{|\theta|>c} |\xi_n(\theta_{n^{-1/2}})|n^{1/2}$ remains bounded as $c \uparrow \infty$; we deduce a contradiction. Let c increase to $+\infty$ through the integers k; let n_k, θ_k, $|\theta_k| \ge k$, achieve the inf when $c = k$, so that $\xi_k(\theta_k n_k^{-1/2})n_k^{1/2}$ is bounded as $k \to \infty$. The identifiability assumption of §3 then implies that $\theta_k/n_k^{1/2} \to 0$ since $\xi_k(\theta_k n_k^{-1/2}) \to 0$. Therefore, since $|\theta_k| \ge k$, $n_k \to +\infty$. Moreover, since $\theta_k n_k^{-1/2}$ goes to zero, the differentiability assumption of §3 applies, so

$$\xi_k(\theta_k n_k^{-1/2}) = \xi_k(0) + \langle\theta_k, \eta\rangle n_k^{-1/2} + o(\theta_k n_k^{-1/2}).$$

Therefore, as $k \to \infty$ with probability approaching 1,

$$|\xi_n(\theta_k n_k^{-1/2})| \ge \tfrac{1}{2}c_0|\theta_k n_k^{-1/2}| - |\xi_n(0)|, \tag{15.9}$$

using nonsingularity ($|\langle\theta, \eta\rangle| \ge c_0|\theta|$). The assumed boundedness of $n_k^{1/2}\xi_n(\theta_k n_k^{-1/2})$ and (15.9) imply that $\{\theta_k\}$ is bounded, a contradiction.

REFERENCES

1. R. J. Beran, *Efficient robust estimation for parametric models*, Z. Wahrsch. Verw. **55** (1981), 91–108.
2. D. Blackwell, *Equivalent comparisons of experiments*, Ann. Math. Statist. **24** (1953), 265–272.
3. E. Bolthausen, *Convergence in distribution of minimum distance estimators*, Metrika **24** (1977), 215–227.
4. Prabir Burman, *Estimation of the mixing distribution*, 1983 (preprint).
5. R. G. Davies, *Asymptotic inference in time series*, Adv. in Appl. Probab. **5** (1973), 469–497.
6. J. L. Doob, *Stochastic processes*, Wiley, 1953.
7. J. Hajek, *Local asymptotic minimax and admissibility in estimation*, Proc. Sixth Berkeley Sympos., Vol. 1, 1972, pp. 170–194.
8. P. J. Huber, *Robust statistics*, Wiley, New York, 1981.
9. A. Ibragimov, *Estimation of the spectral function of a stationary Gaussian process*, Theory Probab. Appl. **8** (1963), 391–430.
10. H. L. Koul and T. de Wet, *Minimum distance estimation in a linear regression model*, Ann. Statist. **11** (1983), 921–932.
11. V. N. La Riccia, *Asymptotic properties of weighted L^2 quantile distance estimators*, Ann. Statist. **10** (1982), 621–624.
12. L. Le Cam, *Sufficiency and asymptotic sufficiency*, Ann. Math. Statist. **35** (1964), 1419–1455.
13. ——, *Théorie asymptotique de la décision statistique*, Presses Univ. Montréal, Montréal, 1969.
14. ——, *Limits of experiments*, Proc. Sixth Berkeley Sympos., Vol. 1, Univ. of California Press, Berkeley, 1972.
15. ——, *Convergence of estimates under dimensionality restrictions*, Ann. Statist. **1** (1973), 38–53.
16. P. W. Millar, *Asymptotic minimax theorems for the sample distribution function*, Z. Wahrsch. Verw. Gebiete **48** (1979), 233–252.
17. ——, *Robust estimation via minimum distance methods*, Z. Wahrsch. Verw. Gebiete **55** (1981), 73–89.
18. W. C. Parr, *Minimum distance estimation: a bibliography*, Comm. Statist. A–Theory Methods **10** (1981), 1205–1224.
19. W. C. Parr and W. R. Schucany, *Minimum distance and robust estimation*, J. Amer. Statist. Assoc. **75** (1980), 616–624.
20. W. C. Parr and T. de Wet, *On minimum weighted Cramer-von Mises statistic estimation*, 1980 (preprint).
21. K. R. Parthasarathy, *Probability measures on metric spaces*, Academic Press, 1967.
22. D. Pollard, *The minimum distance method of testing*, Metrika **27** (1980), 43–70.
23. G. G. Roussas, *Contiguous probability measures: some applications in statistics*, Cambridge Univ. Press, 1972.
24. R. J. Serfling, *Approximation theorems of mathematical statistics*, Wiley, 1980.
25. G. Shorack, *Convergence of quantile and spacings processes with applications*, Ann. Math. Statist. **43** (1972), 1400–1411.
26. J. Wolfowitz, *The minimum distance method*, Ann. Math. Statist. **28** (1957), 75–88.

DEPARTMENT OF STATISTICS, UNIVERSITY OF CALIFORNIA, BERKELEY, CALIFORNIA 94720