Metrika, Volume 27, 1980, page 43–70. ©Physica-Verlag, Vienna.

The Minimum Distance Method of Testing[1])

By *D. Pollard*, New Haven[2])

Abstract: In this paper a method is developed for generalising tests of Kolmogorov-Smirnov and Cramér-von Mises type to cases where parameters have to be estimated. The procedures are based on comparing the empirical distribution function F_n, as a random point in a normed linear space, with a parametric surface $\{F(\underline{\theta}): \underline{\theta} \in \Theta\}$ which represents the family of possible underlying distributions. Asymptotic results are proved for the distribution of the minimum distance $\sqrt{n} \inf_{\underline{\theta}} \| F_n - F(\underline{\theta}) \|$ and for the corresponding minimizing value of $\underline{\theta}$. The results are extended to cases where $\| \cdot \|$ is replaced by a parameter dependent norm $\| \cdot \|_{\underline{\theta}}$, and where the underlying distribution is replaced by a sequence of alternatives. The basic assumptions require convergence in distribution of $\sqrt{n}\,[F_n - F(\underline{\theta}_0)]$ and differentiability in norm of the map $\underline{\theta} \to F(\underline{\theta})$.

1. Introduction

The subject of this paper is the problem of goodness-of-fit testing based on the empirical distribution function, in the case where unknown parameters have to be estimated. The results provide a means for extending the scope of the test procedures associated with the names of Kolmogorov, Smirnov, Cramér and von Mises, i.e. tests based on the asymptotic behaviour of some type of distance between the empirical distribution function F_n and its specified underlying distribution F.

For situations where F depends on an unknown parameter $\underline{\theta}$, a natural procedure would be to form an estimate $\hat{\underline{\theta}}_n$ then compare the distance between F_n and $F(\cdot, \hat{\underline{\theta}}_n)$. Some of the most far-reaching results of this type have been obtained by *Durbin* [1973, 1976], using methods involving convergence in distribution of random elements of $D[0,1]$. Estimating the inverse of the underlying distribution function by $F^{-1}(\cdot, \hat{\underline{\theta}}_n)$, *Durbin* proved that the random functions $\sqrt{n}\,(F_n[F^{-1}(\cdot, \hat{\underline{\theta}}_n)] - \cdot)$ converge in distribution to a Gaussian process with specified covariance structure (which in general depends on the unknown parameter). Goodness-of-fit statistics such as the supremum norm distance between F_n and $F(\cdot, \hat{\underline{\theta}}_n)$ can be recovered by applying various continuous functionals to these random functions.

Alternative measures of fit can be constructed by choosing the estimate of $\underline{\theta}$ to minimize the distance between F_n and $F(\cdot, \underline{\theta})$; this corresponds to using the so-called

---

[1]) Supported by a fellowship of the Alexander von Humbolt Foundation while on leave from Yale University, at the Ruhr-Universität Bochum.

[2]) *David Pollard*, Yale University, Department of Statistics Box 2179 Yale Station, New Haven, Conneticut 06520, U.S.A.

minimum distance estimator $\underline{\theta}_n^*$. Asymptotic results for the distribution of $\underline{\theta}_n^*$ itself have been proved by *Bolthausen* [1977], who chose $\underline{\theta}_n^*$ to minimize $\| F_n[F^{-1}(\cdot, \underline{\theta})] - \cdot \|$ for some suitable norm on $D[0,1]$. For example, use of an L^2 norm leads to statistics of the Cramér-von Mises type, since $\int_0^1 (F_n[F^{-1}(t, \underline{\theta})] - t)^2 dt = \int (F_n(x) - F(x, \underline{\theta}))^2 F(dx, \underline{\theta})$. The reader will recognise a number of extensions of *Bolthausen*'s ideas in the present paper.

The essence of my method consists of reducing the problem to a geometric one of minimizing the distance of a random point in a normed linear space $(X, \| \cdot \|)$ to a prescribed parametrised surface. Instead of a sequence of empirical distribution function in a space such as $D[-\infty, \infty]$, I consider the more general situation of a sequence of random elements $\{F_n\}$ in $(X, \| \cdot \|)$. The role of a parametric family of possible underlying distributions is taken over by a map $\underline{\theta} \to F(\underline{\theta})$ from a subset Θ of $\mathbf{R}^S$ into X. This formulation has the advantage that the same norm $\| \cdot \|$ can serve three different purposes. Firstly, it is involved in the measure of fit $\inf_{\underline{\theta}} \| F_n - F(\underline{\theta}) \|$. Secondly, it enters into the definition of convergence in distribution of the random elements $G_n := \sqrt{n}(F_n - F(\underline{\theta}_0))$. Since X need not be separable (e.g. $D[-\infty, \infty]$ under its sup norm), a slightly modified concept of convergence in distribution must be employed; further details are given in Section 3. Finally, the norm is needed to specify a natural differentiability requirement, viz. that $F(\underline{\theta})$ should be approximable by a linear function $F(\underline{\theta}_0) + \langle \underline{\theta} - \underline{\theta}_0, \underline{D} \rangle$ with an error whose norm is of order $o(|\underline{\theta} - \underline{\theta}_0|)$ near $\underline{\theta}_0$. This concept is discussed in Section 2.

Roughly speaking the method consists of justifying the approximation of $\sqrt{n} \| F_n - F(\underline{\theta}) \|$ for $\underline{\theta}$ near the "true value" $\underline{\theta}_0$ by $\sqrt{n} \| [F(\underline{\theta}_0) + n^{-1/2} G_n] - [F(\underline{\theta}_0) + \langle \underline{\theta} - \underline{\theta}_0, \underline{D} \rangle] \| = \| G_n - \langle \sqrt{n}(\underline{\theta} - \underline{\theta}_0), \underline{D} \rangle \|$. If G_n converges in distribution to some random element G of X, and if the minimum is achieved at a distance of order $O_p(n^{-1/2})$ from $\underline{\theta}_0$, then the distribution of $\sqrt{n} \inf_{\underline{\theta}} \| F_n - F(\underline{\theta}) \|$ should be close to that of $\inf_{\underline{t}} \| G - \langle \underline{t}, \underline{D} \rangle \|$. This heuristic argument is given a rigorous justification in Section 4. Section 5 concerns some slight extensions needed to cover the Cramér-von Mises type of statistics where the distance depends on the parameter $\underline{\theta}$. A method for deriving the asymptotic power under sequences of alternative hypotheses is described in the next section, followed by some of the asymptotic theory for minimum distance estimators. The paper concludes with some comparison with the other method of testing goodness-of-fit mentioned at the start of this introduction.

2. Norm Differentiability

It is my contention that differentiability in norm is the most natural form for the linear approximation property required in the study of minimum distance estimation and testing. Indeed, the regularity conditions to be found in the literature are often only necessary insofar as they imply this property; when X happens to be a space of real functions, norm differentiability can frequently be deduced from various condi-

tions on the existence and regularity of partial derivatives $\partial F/\partial\theta_i$ in the usual sense. The following examples illustrate this point. I find it preferrable, however, to frame the main results in terms of the more general concept, rather than fragment the conditions into a number of assumptions about such partial derivatives.

Recall that $\underline{\theta} \to F(\underline{\theta})$ is a map from $\Theta \subseteq \mathbf{R}^s$ into X. This map is said to be *norm differentiable* at $\underline{\theta}_0$ if there exists a column vector $\underline{D} \in X^s$ such that $\| F(\underline{\theta}) - F(\underline{\theta}_0) - \langle \underline{\theta} - \underline{\theta}_0, \underline{D} \rangle \| = o(| \underline{\theta} - \underline{\theta}_0 |)$ near $\underline{\theta}_0$. Here $\langle \underline{t}, \underline{D} \rangle$ denotes the sum $\sum_i t_i D_i$ where $\underline{t} \in \mathbf{R}^s$ has components $t_1, \ldots, t_s$ and $\underline{D}$ has components $D_1, \ldots, D_s$. Differentiability ensures that $F(\underline{\theta})$ may be approximated by the linear function $F(\underline{\theta}_0) + \langle \underline{\theta} - \underline{\theta}_0, \underline{D} \rangle$ near $\underline{\theta}_0$. In order that the corresponding affine plane in X not be over parametrised, the derivative vector $\underline{D}$ must be *non-singular*, i.e. the components $D_1, \ldots, D_s$ should be linearly independent elements of X. Put another way, if $\underline{t} \neq \underline{0}$ then $\langle \underline{t}, \underline{D} \rangle \neq 0$. Since the function $\underline{t} \to \langle \underline{t}, \underline{D} \rangle$ is continuous and non-zero on the compact set $\{\underline{t} \in \mathbf{R}^s : | \underline{t} | = 1\}$, it follows that nonsingularity of $\underline{D}$ is equivalent to the existence of a constant $C > 0$ for which $\| \langle \underline{t}, \underline{D} \rangle \| \geqslant C | t |$ for all $\underline{t} \in \mathbf{R}^s$. Nonsingularity will be used in this form.

2.1 Example

Consider the location parameter problem as treated by *Blackman* [1955] and *Pyke* [1970]. Here $F(\theta)$ denotes the translation $H(\cdot - \theta)$ of a known distribution function H on $\mathbf{R}$ through a distance specified by the real location parameter θ. A convenient choice for X is the space $D[-\infty, \infty]$ of all real functions on $\mathbf{R}$ which are right continuous with left limits everywhere, and have finite limits at $\pm\infty$. Equip this space with its supremum norm.

For norm differentiability it suffices that H should possess a uniformly continuous density cf. *Pyke* [1970]. Indeed, since $-h(x-\theta_0)$ must converge to zero as $| x | \to \infty$ this function belongs to $D[-\infty, \infty]$ and satisfies

$$\sup_x | H(x-\theta_0-t) - H(x-\theta_0) + th(x-\theta_0) |$$

$$= \sup_x | \int_{x-\theta_0}^{x-\theta_0-t} [h(y) - h(x-\theta_0)]\, dy |$$

$$\leqslant | t | \sup \{ | h(y) - h(y') | : | y - y' | \leqslant | t | \}$$

$$= o(| t |).$$

A similar differentiability property holds for the two parameter scale/location family $H[\gamma(\cdot - \mu)]$; in this case it would suffice to have a uniformly continuous density h for which $h(x) = o(x^{-1})$ as $| x | \to \infty$. □

2.2 Example

In treating the multidimensional case of Cramér-von Mises statistics under parameter estimation, *Neuhaus* [1973] adopted the approach of regarding distribution functions as elements of the Hilbert space $L^2(\mathbf{R}^k, F(d\underline{x}, \underline{\theta}_0))$. The family $\{F(\cdot, \underline{\theta})\}$ of possible underlying distributions was specified by density functions $f(\cdot, \underline{\theta})$ with respect to some σ-finite measure μ. The assumptions made by *Neuhaus* included the requirement that there should exist a neighbourhood U_0 of $\underline{\theta}_0$ in which f had continuous partial derivatives $\partial f/\partial \theta_i$ satisfying the domination condition: there exists a function g_0 with $\int g_0 \, d\mu < \infty$ and $| \partial f/\partial \theta_i | \leqslant g_0$ on U_0. Such a condition guarantees not only differentiability of $F(\cdot, \underline{\theta})$ at $\underline{\theta}_0$ in L^2 norm, but also in the stronger supremum norm sense. Consider for example the case where $\underline{\theta}$ is a two-dimensional parameter.

The vector function $\underline{D}(\underline{x})$ from $\mathbf{R}^k$ into $\mathbf{R}^2$ having components $\int_{(-\infty, \underline{x}]} (\partial f/\partial \theta_i)(\underline{y}, \underline{\theta}_0) \, \mu(d\underline{y})$, for $i = 1, 2$, is an element of the space $[D[-\infty, \infty]^k]^2$ and hence square integrable, and

$$\sup_{\underline{x}} | F(\underline{x}, \underline{\theta}_0 + \underline{t}) - F(\underline{x}, \underline{\theta}_0) - \langle \underline{t}, \underline{D}(\underline{x}) \rangle |$$

$$\leqslant \int | f(\underline{y}, \underline{\theta}_0 + \underline{t}) - f(\underline{y}, \underline{\theta}_0) - t_1 \frac{\partial f}{\partial \theta_1}(\underline{y}, \underline{\theta}_0) - t_2 \frac{\partial f}{\partial \theta_2}(\underline{y}, \underline{\theta}_0) | \, \mu(d\underline{y}).$$

Two applications of the Mean Value Theorem show that the integrand has the form

$$t_1 \left[\frac{\partial f}{\partial \theta_1}(\cdot, \underline{\theta}_0 + \underline{t}^*) - \frac{\partial f}{\partial \theta_1}(\cdot, \underline{\theta}_0) \right] + t_2 \left[\frac{\partial f}{\partial \theta_2}(\cdot, \underline{\theta}_0 + \underline{t}^{**}) - \frac{\partial f}{\partial \theta_2}(\cdot, \underline{\theta}_0) \right]$$

where $| \underline{t}^* |$ and $| \underline{t}^{**} |$ are both less than $| \underline{t} |$. Since the two terms in square brackets are both dominated by $2g_0$, and they converge to zero as $| \underline{t} | \to 0$, it follows that the integral is of order $o(| t |)$, as required.

Notice that the above argument actually shows that $\underline{\theta} \to F(\cdot, \underline{\theta})$ is norm differentiable as a map into $D[-\infty, \infty]^k$ when that space is equipped with its supremum norm. □

2.3 Example

In order to prove very general forms of the functional central limit theorem for empirical measures, *Dudley* [1978] has introduced a class of spaces which incorporates the essential features of spaces such as $D[0,1]$ and $D[-\infty, \infty]^k$. His space $D_0(\mathcal{C}, \lambda)$ can be defined for any class $\mathcal{C}$ of measurable subsets of any probability space $(M, \mathcal{M}, \lambda)$.

First define $C_b(\mathcal{C}, \lambda)$ as the set of all bounded real functions on $\mathcal{C}$ which are continuous with respect to the $L^2(\lambda)$ norm on $\mathcal{C}$. Continuity of a function f on $\mathcal{C}$ in this sense is equivalent to the requirement that $f(C_n) \to f(C)$ whenever $\lambda(C_n \Delta C) \to 0$. For a probability measure P absolutely continuous with respect to λ, *Dudley* [1973; Theorem 2.1] has given conditions ensuring the existence of a version of the so-called "tied down P noise process" G_P having sample paths in $C_b(\mathcal{C}, \lambda)$. This process, which generalises the

notion of a tied down Brownian motion (= Brownian bridge) in $C[0, 1]$, is a Gaussian stochastic process with index set C, with zero mean and having covariance kernel $P(C \cap D) - P(C)\,P(D)$. It often occurs as the limit in distribution of a sequence of normalised empirical measures formed by sampling from the distribution P.

Now define $D_0(C, \lambda)$ as the linear space of real functions on C generated by $C_b(C, \lambda)$ together with the functions $C \to \epsilon_m(C)$, where ϵ_m ranges over all the point masses on M. Equip this space with its supremum norm. Notice that $D_0(C, \lambda)$ contains enough functions to accommodate the sample paths of empirical measures together with the continuous paths of the Gaussian limit processes. The closure of this space under uniform limits corresponds to the usual $D[0,1]$ type of space. As an example, consider the class C of all semi-infinite intervals of the form $(-\infty, \underline{x}]$, including those where some of the coordinates of $\underline{x}$ are $\pm\infty$, in $\mathbf{R}^k$. The classical result on convergence in distribution of the multidimensional empirical distribution function could be formulated in terms of this $D_0(C, P)$. Observe that, if P does not have nonatomic marginal distributions, the space $C_b(C, P)$ in this case includes members corresponding to functions with discontinuities (in the usual sense) at certain fixed points and along certain hyperplanes.

Suppose now that $\{P_{\underline{\theta}}\}$ is a family of probability measures on M dominated by λ. The evaluation $C \to P_{\underline{\theta}}(C)$ defines an element of $C_b(C, \lambda)$ for each $\underline{\theta}$, and hence a map $\underline{\theta} \to F(\underline{\theta})$ from Θ into $C_b(C, \lambda)$. Write $\xi(\underline{\theta})$ for the square root of the density function $f(\underline{\theta}) := dP_{\underline{\theta}}/d\lambda$. I shall show that norm differentiability of $F(\underline{\theta})$ follows from the condition of quadratic differentiability of ξ which was exploited by *Le Cam* [1970] in one of his studies of the asymptotic theory of maximum likelihood estimation. Assume then that there exists an $s \times 1$ column vector $\dot{\underline{\xi}}$ of functions in $L^2(\lambda)$ for which

$$\xi(\underline{\theta}) = \xi(\underline{\theta}_0) + \langle \underline{\theta} - \underline{\theta}_0, \dot{\underline{\xi}} \rangle + R(\underline{\theta} - \underline{\theta}_0) \qquad (*)$$

where $[\int R^2(\underline{\theta} - \underline{\theta}_0)d\lambda]^{1/2} = o(|\,\underline{\theta} - \underline{\theta}_0\,|)$ near $\underline{\theta}_0$.

The function $\underline{D}(C) := \int_C 2\xi(\underline{\theta}_0)\,\dot{\underline{\xi}}\,d\lambda$ certainly defines a column vector of elements of $C_b(C, \lambda)$ since $\xi(\underline{\theta}_0)$ and all the components of $\dot{\underline{\xi}}$ are square integrable with respect to λ, and

$$\sup_{C \in \mathcal{C}} |\,F(\underline{\theta}_0 + \underline{t}) - F(\underline{\theta}_0) - \langle \underline{t}, \underline{D}(C) \rangle\,|$$

$$= \sup_{C \in \mathcal{C}} |\int_C \xi^2(\underline{\theta}_0 + \underline{t}) - \xi^2(\underline{\theta}_0) - 2\xi(\underline{\theta}_0)\langle \underline{t}, \dot{\underline{\xi}} \rangle d\lambda\,|$$

$$\leqslant \int |\langle \underline{t}, \dot{\underline{\xi}} \rangle|^2 + R^2(\underline{t}) + 2\,|\,\xi(\underline{\theta}_0)R(\underline{t})\,| + 2\,|\langle \underline{t}, \dot{\underline{\xi}} \rangle R(\underline{t})\,|\,d\lambda$$

$$= o(|\,\underline{t}\,|)$$

as may be seen by applying the Schwarz inequality to each of the last two terms in the last integrand.

The above result may be formally interpreted as one of "taking a derivative under

the integral sign". For if we write $\dot{\underline{\xi}}$ in the more suggestive form $(\partial/\partial\underline{\theta})f^{1/2}(\underline{\theta}_0)$ and carry out the formal differentiation to give $(1/2)f^{-1/2}(\underline{\theta}_0)\cdot(\partial/\partial\underline{\theta})f(\underline{\theta}_0)$, then we obtain $(\partial/\partial\underline{\theta})\int_C f d\lambda = \int_C (\partial/\partial\underline{\theta}) f d\lambda$ uniformly with respect to C.

Notice also that to check (*) it suffices to verify quadratic differentiability of $\zeta(\underline{\theta}) := (dP_{\underline{\theta}}/d\mu)^{1/2}$ for any dominating measure μ absolutely continuous with respect to λ. For then (*) is satisfied with $\dot{\underline{\xi}} = (d\mu/d\lambda)^{1/2}\cdot\dot{\underline{\zeta}}$. □

3. Convergence in Distribution in Non-Separable Metric Spaces

The space $D[0,1]$ under its supremum norm topology is not separable. Partly because of this fact, we encounter difficulties when trying to analyse the asymptotic behaviour of empirical distributions as random elements in that space: these functions are in general not Borel measurable, cf. *Billingsley* [1968, Section 18]. This problem provided the motivation behind *Dudley*'s [1966, 1967] introduction of a slightly modified concept of convergence in distribution for random elements in a non-separable metric space.

Dudley defined weak convergence for measures defined only on the smaller σ-algebra $\mathcal{B}_0$ generated by the class of all closed balls in any metric space X. For the case of $D[0,1]$ under its supremum norm, this σ-algebra coincides with the cylinder σ-algebra. Empirical distribution functions are therefore $\mathcal{B}_0$ measurable.

The corresponding definition for convergence in distribution of a sequence $\{X_n\}$ of $\mathcal{B}_0$ measurable random elements to a variable X (written $X_n \xrightarrow{\mathcal{D}} X$) consists of two requirements:

(i) the distribution of X concentrates on a separable subset of X;
(ii) $\mathbf{E}f(X_n) \to \mathbf{E}f(X)$ for each bounded, continuous, $\mathcal{B}_0$ measurable, real valued function f on X.

When the topology is separable this definition reduces to the usual one, since $\mathcal{B}_0$ then coincides with the Borel σ-algebra.

Apart from eliminating the measurability problem, this modified concept does not differ all that much from its more well-known counterpart; the two theories run closely parallel cf. *Pollard* [1979a]. The usual results on convergence of the partial sum process to Brownian motion, and of convergence of empirical processes to the Brownian bridge, carry over readily to this setting. The multivariate form of the second of these results even seems more naturally suited to the space $D[-\infty,\infty]^k$ under its supremum norm topology. What is more, we still have at our disposal a form of the Skorokhod representation, a result whose usefulness has been amply demonstrated by *Pyke* [1969, 1970].

3.1 Theorem

Suppose $X_n \xrightarrow{\mathcal{D}} X$ in *Dudley*'s sense. Then there exist versions $\bar{X}_n$ and $\bar{X}$, defined on

some new probability space, for which $\bar{X}_n \overset{\text{a.s.}}{\to} \bar{X}$. □

As usual, "version" means that the new variables induce the same distributions on $\mathcal{B}_0$ as do the original X's. The theorem is a simplified form of a result of *Wichura* [1970].

This theorem is especially well adapted to proving results related to the continuous mapping theorem. The typical sort of argument goes something like this.

Suppose $X_n \overset{\mathcal{D}}{\to} X$ and that $\{h_n\}$ is a sequence of functions converging in some sense to a function h. When does $h_n(X_n) \overset{\mathcal{D}}{\to} h(X)$? Switch to a.s. convergent versions $\bar{X}_n$ and $\bar{X}$. Find conditions on $\{h_n\}$ ensuring that $h_n(\bar{X}_n) \overset{\text{a.s.}}{\to} h(\bar{X})$. This implies the weaker form of convergence $h_n(\bar{X}_n) \overset{\mathcal{D}}{\to} h(\bar{X})$. But $h_n(\bar{X}_n)$ and $h_n(X_n)$ have the same distribution, and so $h(X_n) \overset{\mathcal{D}}{\to} h(X)$ follows. I shall be using arguments of this type several times, without spelling out all the details each time. Once an a.s. convergence statement for one version of the process has been reached, the remainder of the argument will be left to the reader.

In this paper the starting point for such a procedure will generally be a convergence assertion involving an empirical distribution function. This could either be obtained by adapting known results [e.g. Theorem 16.4 of *Billingsley*] to our setting, or by appealing to one of the powerful theorems recently proved by *Dudley* [1978]. For future reference, I shall describe here one of *Dudley*'s results.

Let F_n be the empirical measure obtained by i.i.d. sampling on a probability distribution P. Regard $\sqrt{n}(F_n - P)$ as a random element of the space $D_0(\mathcal{C}, P)$ discussed in Example 2.3. I ignore the question of measurability (with respect to $\mathcal{B}_0$), to which *Dudley* has devoted some effort. Suffice it to say that, for all of cases we shall encounter, *Dudley*'s measurability criteria will prove adequate. A sufficient condition for $\sqrt{n}(F_n - P)$ to converge in distribution to the tied down G_P of Example 2.3 can be framed in terms of the concept of *metric entropy with inclusion*. This quantity depends on the measure P and on the particular class $\mathcal{C} \subseteq \mathcal{M}$ chosen. For each $\epsilon > 0$ define $L(\epsilon)$ as the logarithm of the smallest value of n satisfying: there exist n sets $A_1, \ldots, A_n \in \mathcal{M}$ having the property that to each $C \in \mathcal{C}$ there correspond an A_i and A_j with $A_i \subseteq C \subseteq A_j$ and $P(A_j \setminus A_i) < \epsilon$.

3.2 Theorem

$$\text{If } \int_0^1 L(\epsilon^2)^{1/2} d\epsilon < \infty \text{ then } \sqrt{n}(F_n - P) \overset{\mathcal{D}}{\to} G_P. \qquad \square$$

This is Theorem 5.1 of *Dudley* [1978], but with the question of measurability ignored. The entropy condition here also guarantees the existence of a version of G_P with sample paths in $C_b(\mathcal{C}, P)$; use of this version is understood in the theorem above. It should also be noted that $\sqrt{n}(F_n - P)$ and G_P can be interpreted as random elements in $D_0(\mathcal{C}, \lambda)$, for any probability λ dominating P, and that the convergence in distribution still holds under this interpretation.

By way of illustration, consider the classical situation where $\mathcal{C} = \{(-\infty, x] : -\infty \leqslant x \leqslant \infty\}$ and P is any probability measure on $\mathbf{R}$. To bound $L(\epsilon)$,

determine points $-\infty = x_0 < x_1 < \ldots < x_{m-1} < x_m = +\infty$ inductively by $x_{i+1} := \sup\{x : P(x_i, x] \leqslant \epsilon/2\}$. Since $P(x_i, x_{i+1}] \geqslant \epsilon/2$ we have $m \leqslant 2/\epsilon$. The collection of sets $\{\emptyset, (-\infty, x_1), (-\infty, x_1], (-\infty, x_2), \ldots, (-\infty, x_{m-1}], \mathbf{R}\}$ will serve as our A_i's. Convergence of the entropy integral $\int_0^1 L(\epsilon^2)^{1/2} d\epsilon$ follows from the convergence of $\int_0^1 [\log(4/\epsilon^2)]^{1/2} d\epsilon$. Theorem 3.2 then gives the classical result for convergence in distribution of (normalised) empirical distribution functions on the real line. Extension to the multivariate case can be carried out by techniques similar to those employed by *Elker/Pollard/Stute* [1979] for obtaining the multivariate extension of the classical Glivenko-Cantelli Theorem.

In the theorems to follow the reader will notice that the main requirement regarding F_n is that $\sqrt{n}(F_n - F(\underline{\theta}_0))$ should converge in distribution. In principle then, the results could be applied in various case of dependent sampling where such a central limit result holds. I leave such further developments to the reader.

4. Distribution of the Minimum Distance

Let us first agree on the formulation of the basic model. The following assumptions are to remain in force for the rest of the paper. Given are a normed linear space $(X, \|\cdot\|)$ and a map $\underline{\theta} \to F(\underline{\theta})$ from a subset Θ of $\mathbf{R}^s$ into X. This map will be assumed continuous (and even norm differentiable later). The statistical information comes from a sequence $\{F_n\}$ of random elements of X, defined on a probability space $(\Omega, \mathfrak{F}, \mathbf{P})$, each of which is assumed to be measurable with respect to the σ-algebra $\mathcal{B}_0$ generated by the balls in X. In some sense F_n should converge to $F(\underline{\theta}_0)$ where $\underline{\theta}_0$ is some fixed (but unknown) point in the interior of Θ; if $\underline{\theta}_0$ were a boundary point of Θ then the results would have to be modified along the lines followed by *Chernoff* [1954]. To preserve the analogy with the classical situation I shall sometimes refer to $\underline{\theta}_0$ as the "true value" of $\underline{\theta}$. Our initial concern will be with the limiting distribution of $\sqrt{n} \inf_{\underline{\theta} \in \Theta} \|F_n - F(\underline{\theta})\|$. Notice that, because of the continuity assumption on $F(\underline{\theta})$, there are no problems with measurability here.

As with many problems in asymptotic theory, the argument naturally breaks into two pieces: a global part needed for justifying restricting our attention to values of $\underline{\theta}$ within arbitrarily small neighbourhoods of the true value (cf. consistency); and a local argument, based on the shape of $\|F_n - F(\underline{\theta})\|$ near $\underline{\theta}_0$, which determines the actual form of the limiting distribution. It will prove advantageous to maintain this separation.

The idea behind the global half of the argument is closely related to the one developed by *Wolfowitz* [1957] for proving consistency of minimum distance estimators.

4.1 Lemma

Suppose that $\|F_n - F(\underline{\theta}_0)\| \xrightarrow{\mathbf{P}} 0$ and that, for every neighbourhood N of $\underline{\theta}_0$, $\inf_{\underline{\theta} \notin N} \|F(\underline{\theta}) - F(\underline{\theta}_0)\| > 0$. Then, again for every neighbourhood of $\underline{\theta}_0$,

$$\mathbf{P}\{\inf_{\underline{\theta} \notin N} \| F_n - F(\underline{\theta}) \| > \| F_n - F(\underline{\theta}_0) \|\} \to 1$$

and consequently

$$\mathbf{P}\{\inf_{\underline{\theta} \in \Theta} \| F_n - F(\underline{\theta}) \| = \inf_{\underline{\theta} \in N} \| F_n - F(\underline{\theta}) \|\} \to 1.$$

Proof: From the triangle inequality

$$\| F_n - F(\underline{\theta}) \| \geqslant \| F(\underline{\theta}) - F(\underline{\theta}_0) \| - \| F_n - F(\underline{\theta}_0) \|$$

and therefore

$$\inf_{\underline{\theta} \notin N} \| F_n - F(\underline{\theta}) \| - \| F_n - F(\underline{\theta}_0) \|$$

$$\geqslant \inf_{\underline{\theta} \notin N} \| F(\underline{\theta}) - F(\underline{\theta}_0) \| - 2 \| F_n - F(\underline{\theta}_0) \|$$

$$\xrightarrow{\mathbf{P}} \inf_{\underline{\theta} \notin N} \| F(\underline{\theta}) - F(\underline{\theta}_0) \|$$

$$> 0.$$

The second part follows from the fact that $\underline{\theta}_0 \in N$. □

To verify the separation requirement of this lemma it suffices to show that $F(\underline{\theta}) \neq F(\underline{\theta}_0)$ whenever $\underline{\theta} \neq \underline{\theta}_0$, and that there exists *at least one* compact neighbourhood N_0 of $\underline{\theta}_0$ for which $\inf_{\underline{\theta} \notin N_0} \| F(\underline{\theta}) - F(\underline{\theta}_0) \| > 0$. For then, given an N, the continuous function $\underline{\theta} \to \| F(\underline{\theta}) - F(\underline{\theta}_0) \|$ must be bounded away from zero on the compact set $N_0 \setminus \operatorname{int} N$, and hence also on the set $(\Theta \setminus N_0) \cup (N_0 \setminus \operatorname{int} N) \subseteq \Theta \setminus N$.

Once it has been established that the values of $\underline{\theta}$ determining the asymptotic distribution of the global infimum lie withing a small neighbourhood of $\underline{\theta}_0$, the analysis then depends only upon the form of $F(\underline{\theta})$ near $\underline{\theta}_0$. When the convergence in probability in Lemma 4.1 is strengthened to an assertion concerning the behaviour of the difference $\sqrt{n}\,(F_n - F(\underline{\theta}_0))$, this gives first a result corresponding to the $\sqrt{n}$-consistency of the location of minimising values of $\underline{\theta}$, then the desired asymptotic distribution for the infimum itself.

4.2 Theorem

Suppose the following assumptions hold:

(i) $\inf_{\underline{\theta} \notin N} \| F(\underline{\theta}) - F(\underline{\theta}_0) \| > 0$ for every neighbourhood N of $\underline{\theta}_0$;

(ii) F is norm differentiable with non-singular derivative $\underline{D}$ at $\underline{\theta}_0$;

(iii) there exists a random X valued element G for which

$$G_n := \sqrt{n}\,(F_n - F(\underline{\theta}_0)) \xrightarrow{\mathcal{D}} G,$$

in the sense described in Section 3 for the metric induced by the norm $\| \cdot \|$.

Then the limiting distribution of the goodness-of-fit statistic is given by

$$\sqrt{n} \inf_{\underline{\theta} \in \Theta} \| F_n - F(\underline{\theta}) \| \xrightarrow{\mathcal{D}} \inf_{\underline{t} \in \mathbf{R}^s} \| G - \langle \underline{t}, \underline{D} \rangle \|.$$

Proof

Step I. Because of Lemma 4.1 it suffices to consider only values of $\underline{\theta}$ lying within any particular neighbourhood of $\underline{\theta}_0$. My choice for this neighbourhood will be governed by the remainder term

$$R(\underline{\theta}) := F(\underline{\theta}) - F(\underline{\theta}_0) - \langle \underline{\theta} - \underline{\theta}_0, \underline{D} \rangle.$$

Assumption (ii) can be expressed by saying that there exists an increasing function $\Delta(\epsilon)$ of order $o(1)$ as $\epsilon \downarrow 0$ for which $\| R(\underline{\theta}) \| \leqslant | \underline{\theta} - \underline{\theta}_0 | \cdot \Delta(| \underline{\theta} - \underline{\theta}_0 |)$, and a positive constant C such that $\| \langle \underline{t}, \underline{D} \rangle \| \geqslant C | \underline{t} |$ for all $\underline{t} \in \mathbf{R}^s$. Choose the neighbourhood N_1 of $\underline{\theta}_0$ such that $\Delta(| \underline{\theta} - \underline{\theta}_0 |) \leqslant (1/2) C$ whenever $\underline{\theta} \in N_1$. Now for any value of $\underline{\theta}$,

$$\begin{aligned} \| F_n - F(\underline{\theta}) \| &= \| F_n - F(\underline{\theta}_0) - \langle \underline{\theta} - \underline{\theta}_0, \underline{D} \rangle - R(\underline{\theta}) \| \\ &\geqslant \| \langle \underline{\theta} - \underline{\theta}_0, \underline{D} \rangle \| - \| R(\underline{\theta}) \| - \| F_n - F(\underline{\theta}_0) \|. \end{aligned}$$

Thus for $\underline{\theta} \in N_1$,

$$\| F_n - F(\underline{\theta}) \| - \| F_n - F(\underline{\theta}_0) \| \geqslant \frac{1}{2} C | \theta - \theta_0 | - 2 \| F_n - F(\underline{\theta}_0) \|.$$

Define $\rho_n := 4\sqrt{n} \| F_n - F(\underline{\theta}_0) \| / C$. [By assumption (iii) this random variable converges in distribution to $4 \| G \| / C$, and so it must be of order $O_p(1)$.] This last inequality then implies that the infimum of $\| F_n - F(\underline{\theta}) \|$ over N_1 agrees with its infimum over $N_1 \cap \{\underline{\theta} : \sqrt{n} \, | \underline{\theta} - \underline{\theta}_0 | \leqslant \rho_n\}$. Upon taking Lemma 4.1 into account we deduce that

$$\mathbf{P} \{ \inf_{\underline{\theta} \in \Theta} \| F_n - F(\underline{\theta}) \| = \inf_{\sqrt{n} \, |\underline{\theta} - \underline{\theta}_0| \leqslant \rho_n} \| F_n - F(\underline{\theta}) \| \} \to 1.$$

Step II. In view of the above it makes sense to rescale and work in terms of $\underline{t} = \sqrt{n}(\underline{\theta} - \underline{\theta}_0)$. Define the random set $J_n := \{\underline{t} : | \underline{t} | \leqslant \rho_n \text{ and } \underline{\theta}_0 + \underline{t}/\sqrt{n} \in \Theta\}$. Over this set $\sqrt{n} \| F_n - F(\underline{\theta}) \|$ can be approximated by a simple convex function:

$$\sup_{\underline{t} \in J_n} | \sqrt{n} \| F_n - F(\underline{\theta}_0 + \underline{t}/\sqrt{n}) \| - \| G_n - \langle \underline{t}, \underline{D} \rangle \| |$$

$$= \sup_{\underline{t} \in J_n} | \| \sqrt{n}(F_n - F(\underline{\theta}_0)) - \langle \underline{t}, \underline{D} \rangle - \sqrt{n} R(\underline{\theta}_0 + \underline{t}/\sqrt{n}) \| - \| G_n - \langle \underline{t}, \underline{D} \rangle \| |$$

$$\leqslant \sup_{\underline{t} \in J} \sqrt{n} \| R(\underline{\theta}_0 + \underline{t}/\sqrt{n}\ ) \|$$

$$\leqslant \sup_{\underline{t} \in J_n} \sqrt{n} \cdot | \underline{t}/\sqrt{n} | \cdot \Delta(| \underline{t}/\sqrt{n} |)$$

$$\leqslant \rho_n \Delta(\rho_n / \sqrt{n})$$

$$= o_p(1) \qquad \text{since } \rho_n = O_p(1) \text{ and } \rho_n / \sqrt{n} = o_p(1).$$

Now the continuous convex function $\underline{t} \to \| G_n - \langle \underline{t}, \underline{D} \rangle \|$ achieves its overall minimum at a point in $\{\underline{t}: | \underline{t} | \leqslant \rho_n\}$; for if $| \underline{t} | > \rho_n$ then

$$\| G_n - \langle \underline{t}, \underline{D} \rangle \| \geqslant C | \underline{t} | - \| G_n \|$$

$$> 3 \| G_n \|$$

$$\geqslant \| G_n - \langle \underline{0}, \underline{D} \rangle \|.$$

With probability tending to one, this minimizing value lies in J_n because $\underline{\theta}_0 \in \text{int } \Theta$. We thus have two functions which, with high probability, are uniformly close over J_n and whose unrestricted infima are the same as their infima over J_n. This implies that

$$\sqrt{n} \inf_{\underline{\theta} \in \Theta} \| F_n - F(\underline{\theta}) \| = \inf_{\underline{t} \in \mathbf{R}^s} \| G_n - \langle \underline{t}, \underline{D} \rangle \| + o_p(1).$$

Step III. To prove that $\inf_{\underline{t} \in \mathbf{R}^s} \| G_n - \langle \underline{t}, \underline{D} \rangle \| \xrightarrow{\mathcal{D}} \inf_{\underline{t} \in \mathbf{R}^s} \| G - \langle \underline{t}, \underline{D} \rangle \|$ we have only to apply the continuous mapping theorem for the functional $m(x) := \inf_{\underline{t} \in \mathbf{R}^s} \| x - \langle \underline{t}, \underline{D} \rangle \|$.

This functional is both $\mathcal{B}_0$ measurable (take the infimum over a countable dense set of $\underline{t}$ values) and continuous (since $| m(x) - m(y) | \leqslant \| x - y \|$). [The continuous mapping theorem in this simple form follows directly from the definition of convergence in distribution given in Section 3.] □

From this theorem we can deduce results of the Kolmogorov-Smirnov type without much trouble. More importantly, though, the method of proof will serve as the model for the various extensions to be presented in Sections 5, 6 and 7.

4.3 Example

Consider once more the location problem described in Example 2.1. Take F_n as the empirical distribution function based on random sampling from $H(\cdot - \theta_0)$. As discussed in Section 3, the process $\sqrt{n}\,[F_n - H(\cdot - \theta_0)]$ can be shown to converge in distribution, as a random element of $D[-\infty, \infty]$, to a Gaussian process. This limit process

may be identified with $W^0(H(\cdot - \theta_0))$, where W^0 denotes the usual Brownian bridge on [0,1] [see Theorem 16.4 of *Billingsley*].

If H has a uniformly continuous density h then, as was shown in Example 2.1, the function $-h(\cdot - \theta_0)$ plays the role of the derivative $\underline{D}$. The parameter space $\mathbf{R}$ being one dimensional, no problems with singularity of this derivative can arise. Notice also that $\liminf_{|\theta| \to \infty} \| H(\cdot - \theta) - H(\cdot - \theta_0) \| > 0$, which shows that the separation property holds.

It follows therefore that

$$\sqrt{n} \inf_{\theta} \| F_n - H(\cdot - \theta) \| \xrightarrow{\mathcal{D}} \inf_{t} \| W^0(H(\cdot - \theta_0)) + th(\cdot - \theta_0) \|.$$

The right hand side can be further simplified to $\inf \| W^0(H(\cdot)) + th(\cdot) \|$, a variable whose distribution does not depend on the unknown θ_0. □

4.4 Example

Using the results of Example 2.3 and Theorem 3.2 (or one of the other theorems proved by *Dudley*, 1978) a general asymptotic result for statistics of the form $\sqrt{n} \inf_{\underline{\theta}} \sup_{C \in \mathcal{C}} | F_n(C) - P_{\underline{\theta}}(C) |$ could be obtained. Even for the relatively simple cases where $\mathcal{C}$ consists of intervals $(-\infty, \underline{x}]$ in $\mathbf{R}^k$ though, several difficulties remain which make the results far from satisfactory. Firstly it appears difficult to give general criteria ensuring non-singularity of the derivative $\underline{D}$ – this is a common problem with theorems of this type. Also, the separation assumption would have to be checked by means of some special features of the class $\mathcal{C}$ and the family $\{P_{\underline{\theta}}\}$. Worst of all, the limit distribution will depend on the unknown parameter $\underline{\theta}_0$ in a possibly complicated fashion. If this dependence on $\underline{\theta}_0$ were continuous then there would certainly be asymptotic procedures available for constructing tests of fit, cf. *Chernoff/Lehmann* [1954], *Neuhaus* [1973] or *Csörgő* et al. [1974]; but these procedures would involve a prohibitive amount of calculation. □

5. Minimization with Parameter Dependent Norms

The methods of the previous section can be modified to cover situations where the distance between F_n and $F(\underline{\theta})$ is measured by a norm $\| \cdot \|_{\underline{\theta}}$ depending on $\underline{\theta}$, and it is the quantity $\| F_n - F(\underline{\theta}) \|_{\underline{\theta}}$ which should be minimized. Two procedures of this type come to mind: Cramér-von Mises tests where the measure of fit is taken as $\int [F_n(x) - F(x, \underline{\theta})]^2 F(dx, \underline{\theta})$; and the method of minimum χ^2 in the classical χ^2 goodness-of-fit test. The local parts of the argument for both these procedures can be handled by a generalised form of Theorem 4.2; the global aspects seem to require two distinct, but clearly related, approaches. An alternative method, based on the use of a preliminary consistent estimate for $\underline{\theta}_0$, is given by Theorem 5.6.

The basic setting of random elements F_n in a normed linear space $(X, \| \cdot \|)$ and a

continuous map $\underline{\theta} \to F(\underline{\theta})$ from Θ into X etc. as described in Section 4, remains the same. In addition X will be equipped with a parametric family $\{\| \cdot \|_{\underline{\theta}} : \underline{\theta} \in \Theta\}$ of norms which must satisfy the following regularity conditions.

5.1 Assumptions on $\| \cdot \|_{\underline{\theta}}$

(i) there exists a neighbourhood N^* of $\underline{\theta}_0$ for which $\| x \|_{\underline{\theta}} \leqslant \| x \|$ for all $x \in X$ and all $\underline{\theta} \in N^*$;
(i) for each fixed $\underline{\theta}$, the map $x \to \| x \|_{\underline{\theta}}$ is $\mathcal{B}_0$ measurable;
(iii) for each fixed x, the map $\underline{\theta} \to \| x \|_{\underline{\theta}}$ is continuous;
(iv) the map $x \to \inf_{\underline{\theta} \in \Theta} \| x - F(\underline{\theta}) \|_{\underline{\theta}}$ is $\mathcal{B}_0$ measurable.

In view of (ii), assumption (iv) would be redundant if the infimum could be replaced by an infimum over a countable dense subset of Θ. Because of the continuity of $\underline{\theta} \to F(\underline{\theta})$, such would be the case if the map $(x, \underline{\theta}) \to \| x \|_{\underline{\theta}}$ were continuous. Assumptions (i) and (iii) in fact imply the slightly weaker property, that this map is continuous on $X \times N^*$, since $| \| x \|_{\underline{\theta}} - \| x' \|_{\underline{\theta}'} | \leqslant | \| x \|_{\underline{\theta}} - \| x \|_{\underline{\theta}'} | + \| x - x' \|$ when $\underline{\theta}' \in N^*$. This ensures $\mathcal{B}_0$ measurability of quantities like $\inf_{\underline{\theta} \in N} \| x - F(\underline{\theta}) \|_{\underline{\theta}}$ whenever $N \subseteq N^*$.

5.2 Lemma

Suppose that $\| F_n - F(\underline{\theta}_0) \| \xrightarrow{\mathbf{P}} 0$. Then

$$\mathbf{P}_* \{ \inf_{\underline{\theta} \notin N} \| F_n - F(\underline{\theta}) \|_{\underline{\theta}} > \| F_n - F(\underline{\theta}_0) \|_{\underline{\theta}_0} \} \to 1.$$

under either of the following assumptions:

a) for each neighourhood N of $\underline{\theta}_0$ there exists a constant $k > 0$ such that $k \| x \| \leqslant \| x \|_{\underline{\theta}}$ for all x whenever $\underline{\theta} \notin N$; also $\inf_{\underline{\theta} \notin N} \| F(\underline{\theta}) - F(\underline{\theta}_0) \| > 0$;

b) for each neighourhood N of $\underline{\theta}_0$ there exists a constant $k > 0$ such that $k \| x \| \geqslant \| x \|_{\underline{\theta}}$ for all x whenever $\underline{\theta} \notin N$; also $\inf_{\underline{\theta} \notin N} \| F(\underline{\theta}) - F\underline{\theta}_0) \|_{\underline{\theta}} > 0$.

Proof: The argument in both cases is similar to that in Lemma 4.1. For part a) use the inequality

$$\inf_{\underline{\theta} \notin N} \| F_n - F(\underline{\theta}) \|_{\underline{\theta}} - \| F_n - F(\underline{\theta}_0) \|_{\underline{\theta}_0}$$
$$\geqslant k \cdot \inf_{\underline{\theta} \notin N} \| F(\underline{\theta}) - F(\underline{\theta}_0) \| - (k+1) \| F_n - F(\underline{\theta}_0) \|,$$

and for part b)

$$\inf_{\underline{\theta} \notin N} \| F_n - F(\underline{\theta}) \|_{\underline{\theta}} - \| F_n - F(\underline{\theta}_0) \|_{\underline{\theta}_0}$$
$$\geqslant \inf_{\underline{\theta} \notin N} \| F(\underline{\theta}) - F(\underline{\theta}_0) \|_{\underline{\theta}} - (k+1) \| F_n - F(\underline{\theta}_0) \|.$$

□

Once again the asymptotic distribution of the minimum distance can be obtained by strengthening the convergence in probability to convergence in distribution of $\sqrt{n}\,(F_n - F(\underline{\theta}_0))$, and introducing the norm differentiability (with non-singular derivative) condition for $F(\underline{\theta})$ at $\underline{\theta}_0$. Both of these requirements are to be interpreted as for Theorem 4.2, i.e. in terms of the norm $\| \cdot \|$. This time though, non-singularity of the derivative $\underline{D}$ can be used to prove the existence of a positive constant C such that $\| \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}} \geqslant C \mid \underline{t} \mid$ for all $\underline{t} \in \mathbf{R}^s$ and all $\underline{\theta}$ near enough to $\underline{\theta}_0$. For, without loss of generality, we may suppose that the neighbourhood N^* of Assumption 5.1 (i) is compact. Continuity of $(x, \underline{\theta}) \to \| x \|_{\underline{\theta}}$ on $X \times N^*$ implies continuity of the map $(\underline{t}, \underline{\theta}) \to \| \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}}$ on $\mathbf{R}^s \times N^*$; this function must therefore be bounded away from zero on the compact set $\{\underline{t} \in \mathbf{R}^s : \mid \underline{t} \mid = 1\} \times N^*$. We can choose C as the lower bound over this set.

5.3 Theorem

Suppose that

(i) at least one of the assumptions a) and b) of Lemma 5.2 holds;
(ii) F is norm differentiable with non-singular derivative $\underline{D}$ at $\underline{\theta}_0$;
(iii) there exists a random X valued element G for which $G_n := \sqrt{n}\,(F_n - F(\underline{\theta}_0)) \xrightarrow{\mathcal{D}} G$.

The the limiting distribution of the goodness-of-fit statistics is given by

$$\sqrt{n} \inf_{\underline{\theta} \in \Theta} \| F_n - F(\underline{\theta}) \|_{\underline{\theta}} \xrightarrow{\mathcal{D}} \inf_{\underline{t} \in \mathbf{R}^s} \| G - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}_0}.$$

Proof

Step I. Use the same definition of $R(\underline{\theta})$ and $\Delta(\epsilon)$ as in Step I of Theorem 4.2, but this time take C as the constant defined in the discussion above. Then choose the neighbourhood $N_1 \subseteq N^*$ so that $\Delta(\mid \underline{\theta} - \underline{\theta}_0 \mid) \leqslant (1/2)\, C$ on N_1. As before we have

$$\| F_n - F(\underline{\theta}) \|_{\underline{\theta}} \geqslant \| \langle \underline{\theta} - \underline{\theta}_0, \underline{D} \rangle \|_{\underline{\theta}} - \| R(\underline{\theta}) \|_{\underline{\theta}} - \| F_n - F(\underline{\theta}_0) \|_{\underline{\theta}},$$

and thus for $\underline{\theta} \in N_1$,

$$\| F_n - F(\underline{\theta}) \|_{\underline{\theta}} - \| F_n - F(\underline{\theta}_0) \|_{\underline{\theta}_0} \geqslant (1/2)\, C \mid \underline{\theta} - \underline{\theta}_0 \mid - 2 \| F_n - F(\underline{\theta}_0) \|.$$

Defining $\rho_n := 4 \| G_n \| / C$ as in Theorem 4.2, we then conclude that the search for the infimum of $\| F_n - F(\underline{\theta}) \|_{\underline{\theta}}$ may be restricted to values of $\underline{\theta}$ of the form $\underline{\theta}_0 + \underline{t}/\sqrt{n}$, where $\underline{t} \in J_n := \{\underline{t} : \mid \underline{t} \mid \leqslant \rho_n$ and $\underline{\theta}_0 + \underline{t}/\sqrt{n} \in N^*\}$.

Step II. The same argument as before shows that

$$\sup_{\underline{t} \in J_n} \mid \sqrt{n} \| F_n - F(\underline{\theta}_0 + \underline{t}/\sqrt{n}) \|_{\underline{\theta}_0 + \underline{t}/\sqrt{n}} - \| G_n - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}_0 + \underline{t}/\sqrt{n}} \mid$$

is of order $o_p(1)$, and therefore that

$$\sqrt{n} \inf_{\underline{\theta} \in \Theta} \| F_n - F(\underline{\theta}) \|_{\underline{\theta}} = \inf_{\underline{t} \in J_n} \| G_n - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}_0 + \underline{t}/\sqrt{n}} + o_p(1).$$

We cannot replace the infimum on the right hand side by an unrestricted infimum at this stage, because of the presence of the $\underline{t}$ in $\| \cdot \|_{\underline{\theta}_0 + \underline{t}/\sqrt{n}}$. However this small complication will disappear in the limit.

Step III. This time a slightly more complicated form of the continuous mapping theorem will be required. The easiest approach is via the use of a.s. convergent versions together with a convergence theorem for convex functions. By Theorem 3.1 there exist versions of the G_n and G processes for which $\bar{G}_n \overset{\text{a.s.}}{\to} \bar{G}$. All of the preceding arguments still apply when ρ_n is replaced by its corresponding $\bar{\rho}_n$, and J_n by $\bar{J}_n$ etc. As explained in Section 3, it suffices to prove an a.s. convergence result for the new process, in order to obtain the distribution convergence result for the original G_n's. With this in mind, select and fix a point $\bar{\omega}$ of the new underlying probability space, a point at which $\| \bar{G}_n(\bar{\omega}) - \bar{G}(\bar{\omega}) \| \to 0$.

The first thing to notice is that

$$\sup_{t \in \mathbf{R}} \sup_{\underline{\theta} \in N^*} | \| \bar{G}_n(\bar{\omega}) - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}} - \| \bar{G}(\bar{\omega}) - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}} | \leqslant \| \bar{G}_n(\bar{\omega}) - \bar{G}(\bar{\omega}) \|.$$

Thus it suffices to show that

$$\inf_{\underline{t} \in \bar{J}_n(\bar{\omega})} \| \bar{G}(\bar{\omega}) - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}_0 + \underline{t}/\sqrt{n}} \to \inf_{\underline{t} \in \mathbf{R}^s} \| \bar{G}(\bar{\omega}) - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}_0}.$$

The functions involved in the lefthand side of this last expression need not be convex in $\underline{t}$, because of the presence of the $\underline{\theta}_0 + \underline{t}/\sqrt{n}$ on the norm. To avoid this problem it will pay to introduce the artifice of a second variable and define functions

$$g_n(\underline{s}, \underline{t}) := \| \bar{G}(\bar{\omega}) - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}_0 + \underline{s}/\sqrt{n}}$$

and

$$g(\underline{t}) \quad := \| \bar{G}(\bar{\omega}) - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}_0}.$$

For each fixed $\underline{s}$, the functions $g_n(\underline{s}, \cdot)$ are convex. Now choose a compact set $K \subseteq \mathbf{R}^s$ containing all the $\bar{J}_n(\bar{\omega})$'s; this is possible since $\bar{\rho}_n(\bar{\omega}) \to 4 \| \bar{G}(\bar{\omega}) \| / C$. The continuity assumption 5.1 (iii) implies that, for each fixed $x \in X$, $\sup_{\underline{s} \in K} | \| x \|_{\underline{\theta}_0 + \underline{s}/\sqrt{n}} - \| x \|_{\underline{\theta}_0} | \to 0$. Taking $x = \bar{G}(\bar{\omega}) - \langle \underline{t}, \underline{D} \rangle$ we thus obtain the result that, for each fixed $\underline{t}$, the sequence $g_n(\underline{s}, \underline{t})$ converges uniformly in $\underline{s} \in K$ to $g(\underline{t})$. In particular, the family $\{g_n(\underline{s}, \cdot): \underline{s} \in K, n = 1, 2, \ldots\}$ of convex functions is pointwise bounded. Theorem 10.6 of *Rockafellar* [1972] thus guarantees the existence

of a real number λ such that

$$|g_n(\underline{s}, \underline{t}) - g_n(\underline{s}, \underline{t}')| \leqslant \lambda |\underline{t} - \underline{t}'|$$

for every n, every $\underline{s} \in K$ and every pair of point $\underline{t}$ and $\underline{t}'$ in K. With all of these uniformity properties we can hardly help but get uniform convergence of $g_n(\underline{s}, \underline{t})$ to $g(\underline{t})$ over $K \times K$.

Choose a finite subset K_0 of K with the property that for each $\underline{t} \in K$ there exists a $\underline{t}^* \in K_0$ for which $\lambda |\underline{t} - \underline{t}^*| < \epsilon$ and $|g(\underline{t}) - g(\underline{t}^*)| < \epsilon$. Then

$$\sup_{\underline{s}, \underline{t} \in K} |g_n(\underline{s}, \underline{t}) - g(\underline{t})|$$

$$\leqslant \sup_{\underline{s}, \underline{t} \in K} |g_n(\underline{s}, \underline{t}) - g_n(\underline{s}, \underline{t}^*)| + \sup_{\underline{s} \in K, \underline{t}^* \in K_0} |g_n(\underline{s}, \underline{t}^*) - g(\underline{t}^*)| +$$

$$+ \sup_{\underline{t} \in K} |g(\underline{t}^*) - g(\underline{t})|$$

from which the desired uniform convergence follows. In particular,

$$\sup_{\underline{t} \in K} \left| \| \bar{G}(\bar{\omega}) - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}_0 + \underline{t}/\sqrt{n}} - \| \bar{G}(\bar{\omega}) - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}_0} \right| \to 0,$$

which implies that

$$\inf_{\underline{t} \in \bar{J}_n(\bar{\omega})} \| \bar{G}(\bar{\omega}) - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}_0 + \underline{t}/\sqrt{n}} - \inf_{\underline{t} \in \bar{J}_n(\bar{\omega})} \| \bar{G}(\bar{\omega}) - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}_0} \to 0.$$

The infimum in the second term can be replaced by an unrestricted infimum because $\| \bar{G}(\bar{\omega}) - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}_0}$ certainly attains its overall infimum somewhere in the region $\{\underline{t}: |\underline{t}| \leqslant 2 \| \bar{G}(\bar{\omega}) \| / C$, a set which is eventually contained in $\bar{J}_n(\bar{\omega})$. [Notice that for $|t| > 2 \| \bar{G}(\bar{\omega}) \| / C$,

$$\| \bar{G}(\bar{\omega}) - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}_0} - \| \bar{G}(\bar{\omega}) - \langle \underline{0}, \underline{D} \rangle \|_{\underline{\theta}_0}$$

$$\geqslant C|t| - 2 \| \bar{G}(\bar{\omega}) \|_{\underline{\theta}_0}$$

$$> 0].$$ □

Since all of the points of interest in the application of this theorem to Cramér-von Mises type statistics can be brought out through the simple location parameter example, I shall restrict myself to this case. The reader will assuredly be able to piece together the appropriate norm differentiability and convergence in distribution results to obtain the details for the more general cases.

5.4 Example

Consider yet again the location problem discussed in Example 2.1 and 4.3. This time the minimization involves $\| F_n - F(\theta) \|_\theta^2 := \int [F_n(x) - F(x, \theta)]^2 F(dx, \theta)$, where $F(\cdot, \theta)$ denotes the translate $H(\cdot - \theta)$. For the normed linear space $(X, \| \cdot \|)$ we have a choice. Perhaps the simpler possibility is $D[-\infty, \infty]$ under its supremum norm, but it would also be possible to equip the same space with the (pseudo) norm $\| \cdot \| := \sup_\theta \| \cdot \|_\theta$. The latter would have the advantage of leading to a separable topology. Under the condition on h used before, all of the required assumptions are easily checked. The corresponding limit result then takes the form

$$n \inf_\theta \int [F_n(x + \theta) - H(x)]^2 H(dx) \xrightarrow{\mathcal{D}} \inf_t \int [W^o(H(x)) + th(x)]^2 H(dx).$$

The limit distribution again does not depend on the unknown parameter θ_0. The expression on the right hand side can be put in a more explicit form, since the function to be minimized reduces with a change of variable to the quadratic in t:

$$\int_0^1 W^o(y)^2 dy + 2t \int_0^1 W^o(y) h(H^{-1}(y)) dy + t^2 \int_0^1 h[H^{-1}(y)]^2 dy$$

whose minimum equals

$$\int_0^1 W^o(y)^2 dy - [\int_0^1 W^o(y) h(H^{-1}(y))\, dy]^2 / \int_0^1 h[H^{-1}(y)]^2 dy.$$

Approximations to the distribution of such quadratic forms involving the Brownian bridge W^o may be obtained by eigenfunction expansions into a series of weighted non-central χ^2 variates; the reader is referred to the papers of *Darling* [1955], *Kac/Kiefer/Wolfowitz* [1955], *Neuhaus* [1973] plus other references given by *Neuhaus* [1977].

One slight problem has been overlooked so far, viz. the $\| \cdot \|_\theta$'s might only be pseudonorms. The only difficulty this could cause would be in the argument leading up to the existence of the constant C. There it was necessary that $\| \langle \underline{t}, \underline{D} \rangle \|_\theta \neq 0$ whenever $| \underline{t} | = 1$ and θ is near θ_0. Convexity of the functions $\underline{t} \to \| \langle \underline{t}, \underline{D} \rangle \|_\theta$ ensures however that $\sup_{|\underline{t}|=1} | \| \langle \underline{t}, \underline{D} \rangle \|_\theta - \| \langle \underline{t}, \underline{D} \rangle \|_{\theta_0} | \to 0$ as $\theta \to \theta_0$. Thus it would suffice to interpret non-singularity of $\underline{D}$ as meaning that $\| \langle \underline{t}, \underline{D} \rangle \|_{\theta_0} \neq 0$ whenever $\underline{t} \neq \underline{0}$. For the problem at hand, non-singularity in this sense is easily checked. □

When application of the general result of Theorem 5.3 to any specific problem is envisaged, it should not prove surprising if special features of that particular problem should render parts of the general proof redundant. This is indeed the case with application to the method of minimum χ^2, where much of the argument in Step III could be avoided because of finite dimensionality considerations. Nevertheless, the proof as it stands still compares favourably with other known methods of approach – see for example Section 2.7 of *Witting/Nölle* [1970]. What is more, as I shall show in a future

paper [*Pollard*, 1979b], the methods developed above lead to a simple analysis for the problem of χ^2 goodness-of-fit testing (using the maximum likelihood estimator) with random cells.

5.5 Example

Observations are taken on a multinomial distribution over k cells, where the probability $p_i(\underline{\theta})$ of falling into the i-th cell depends on an unknown parameter $\underline{\theta}$. The set Θ of possible values for $\underline{\theta}$ is a subset of $\mathbf{R}^s$ which contains the true value $\underline{\theta}_0$ in its interior. Arrange $p_1(\underline{\theta}), \ldots, p_k(\underline{\theta})$ into a column vector $F(\underline{\theta})$. Taking X as $\mathbf{R}^k$, we thus have our map $\underline{\theta} \to F(\underline{\theta})$ from Θ into X. For the i-th component of the random column vector F_n take the proportion of the first n observations which fall into cell i.

With each $\underline{\theta}$ associate a diagonal matrix $\Lambda(\underline{\theta}) := \operatorname{diag}(p_1(\underline{\theta}), \ldots, p_k(\underline{\theta}))$ and a norm $\| \underline{x} \|_{\underline{\theta}} := (\underline{x}'\Lambda^{-1}(\underline{\theta})\underline{x})^{1/2}$ on $\mathbf{R}^k$. We must of course assume all the $p_i(\underline{\theta})$'s to be positive. As the norm $\| \cdot \|$ on X take a multiple of the usual Euclidean norm large enough so that $\| \cdot \| \geqslant \| \cdot \|_{\underline{\theta}}$ for all $\underline{\theta}$ close enough to $\underline{\theta}_0$.

The test of goodness-of-fit derived from the method of minimum χ^2 consists of minimizing the quantity $X_n^2(\underline{\theta}) := n \| F_n - F(\underline{\theta}) \|_{\underline{\theta}}^2$ over Θ. Under appropriate conditions this mimimum has a limiting χ^2_{k-s-1} distribution. The simplest set of conditions seems to be those used by *Birch* [1964] for obtaining the limiting distribution of $X_n^2(\hat{\underline{\theta}}_n)$, where $\hat{\underline{\theta}}_n$ is an estimator of maximum likelihood type. These conditions are essentially equivalent to:

(i) for each neighbourhood N of $\underline{\theta}_0$

$$\inf_{\underline{\theta} \notin N} \| F(\underline{\theta}) - F(\underline{\theta}_0) \| > 0;$$

(ii) there exists a non-singular $k \times s$ matrix D such that

$$F(\underline{\theta}) = F(\underline{\theta}_0) + D(\underline{\theta} - \underline{\theta}_0) + o(|\underline{\theta} - \underline{\theta}_0|) \text{ near } \underline{\theta}_0.$$

Since each of the $p_i(\underline{\theta})$'s can be not greater than one, it is easy to verify condition a) of Lemma 5.2. Continuity of each $p_i(\underline{\theta})$ would take care of all the assumptions on $\| \cdot \|_{\underline{\theta}}$, although the result still holds under somewhat weaker assumptions in this case. The ordinary multivariate form of the central limit theorem ensures that $\sqrt{n}(F_n - F(\underline{\theta}_0)) \xrightarrow{D} \underline{Z}$, where $\underline{Z}$ has a $N(\underline{0}, \Lambda(\underline{\theta}_0) - F(\underline{\theta}_0)F(\underline{\theta}_0)')$ distribution. From Theorem 5.3 it follows under these assumptions that

$$\inf_{\underline{\theta}} X_n^2(\underline{\theta}) \to \inf_{\underline{t}} \| \underline{Z} - D\underline{t} \|_{\underline{\theta}_0}^2.$$

Standard techniques of multivariate normal distribution theory show that the right hand side of this last expression has the desired χ^2_{k-s-1} distribution. □

As evidenced by the need for Lemma 5.2, the global part of the argument for Theo-

rem 5.3 has a slightly ad hoc flavour to it. This was caused by the possibility of strange behaviour of the norm $\| \cdot \|_{\underline{\theta}}$ for values of $\underline{\theta}$ a long way from $\underline{\theta}_0$. To some extent this difficulty can be overcome by using a preliminary consistent estimate $\hat{\underline{\theta}}_n$ of $\underline{\theta}_0$ in the norm $\| \cdot \|_{\hat{\underline{\theta}}_n}$ measuring the goodness-of-fit. Such consistent estimates are generally regarded as being easily obtainable.

5.6 Theorem

Suppose that

(i) there exists a neighbourhood N' of $\underline{\theta}_0$ such that, for every other neighbourhood N of $\underline{\theta}_0$, we have $\inf_{\underline{\theta} \notin N} \inf_{\underline{\theta}' \in N'} \| F(\underline{\theta}) - F(\underline{\theta}_0) \|_{\underline{\theta}'} > 0$;

(ii) F is norm differentiable with non-singular derivative $\underline{D}$ at $\underline{\theta}_0$;

(iii) $G_n := \sqrt{n}\,(F_n - F(\underline{\theta}_0)) \xrightarrow{\mathcal{D}} G$.

Then for any sequence of estimators $\{\hat{\underline{\theta}}_n\}$ with $\hat{\underline{\theta}}_n \xrightarrow{\mathbf{P}} \underline{\theta}_0$,

$$\sqrt{n} \inf_{\underline{\theta} \in \Theta} \| F_n - F(\underline{\theta}) \|_{\hat{\underline{\theta}}_n} \xrightarrow{\mathcal{D}} \inf_{\underline{t} \in \mathbf{R}^s} \| G - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}_0}.$$

Proof: By almost the same procedure as before we arrive at the stage where

$$\sqrt{n} \inf_{\underline{\theta} \in \Theta} \| F_n - F(\underline{\theta}) \|_{\hat{\underline{\theta}}_n} = \inf_{\underline{t} \in \mathbf{R}^s} \| G_n - \langle \underline{t}, \underline{D} \rangle \|_{\hat{\underline{\theta}}_n} + o_p(1).$$

This time there are no difficulties about taking the unrestricted infimum on the right hand side since, with probability tending to one, $\hat{\underline{\theta}}_n$ belongs to the neighbourhood N^* of $\underline{\theta}_0$ in which the inequality $\| \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}} \geqslant C \mid \underline{t} \mid$ holds for all $\underline{t}$.

For the analogue of Step III we need to consider the limiting behaviour of $(G_n, \hat{\underline{\theta}}_n)$ as a random element of the space $X \times \mathbf{R}^s$. Equip this space with the norm $\| (x, \underline{\theta}) \| := \max \{ \| x \|, \mid \underline{\theta} \mid \}$. Since the σ-algebra generated by the balls for this norm coincides with the product of $\mathcal{B}_0$ with the Borel σ-algebra on $\mathbf{R}^s$, no problems with measurability arise. Very slight modifications of the standard argument [Theorem 4.4 of *Billingsley*] then show that $(G_n, \hat{\underline{\theta}}_n) \xrightarrow{\mathcal{D}} (G, \underline{\theta}_0)$. Change over to a.s. convergent versions $(\bar{G}_n, \bar{\underline{\theta}}_n) \xrightarrow{\text{a.s.}} (\bar{G}, \underline{\theta}_0)$. The rest of the argument is then similar to that of Theorem 5.3. [Theorem 10.8 of *Rockafellar*, 1972 can be used to demonstrate uniform convergence on compact $\underline{t}$ sets of $\| \bar{G}_n(\bar{\omega}) - \langle \underline{t}, \underline{D} \rangle \|_{\bar{\underline{\theta}}_n(\bar{\omega})}$ to $\| \bar{G}(\bar{\omega}) - \langle \underline{t}, \underline{D} \rangle \|_{\underline{\theta}_0}$.] □

The application of this result to the problems of Examples 5.4 and 5.5. is straightforward. Further refinements along the same lines are possible, but I don't wish to labour this point. For example, for Cramér-von-Mises statistics the squared "random norm" $\int [\,\cdot\,]^2 F(dx, \hat{\underline{\theta}}_n)$ could be replaced by $\int [\,\cdot\,]^2 F_n(dx)$ under appropriate circumstances.

6. Power of the Tests Under Alternatives

For a test to be considered as a useable procedure, there should at least be some method for gauging its power under an alternative hypothesis. When treated as an asymptotic result for tests of the type considered in this paper, this would lead to the rather uninformative statement to the effect that the power against any fixed alternative tends to one as the sample size increases. In order to get around this problem it has become customary to consider not just a fixed alternative, but rather a sequence of alternatives approaching the null hypothesis at a rate which produces a non-trivial limiting value for the power. This then gives some measure of the discriminating power (against close alternatives) of the proposed test.

In the framework adopted in this paper such an approach can be handled without changing the basic setting; the effect of certain sequences of alternatives can be completely specified by modifying the limit in distribution of $\sqrt{n}\,(F_n - F(\underline{\theta}_0))$. At least the parametric alternatives of *Durbin* [1973] fall into this category. To illustrate the procedure in a setting which incorporates most of the essential features, I return to the situation considered in Example 2.3.

Instead of supposing that F_n denotes the empirical distribution function obtained by a sample of size n from a fixed $P_{\underline{\theta}_0}$, assume that the underlying distribution actually depends on the value of n i.e. F_n represents the empirical distribution function from n independent observations on a probability distribution Q_n. The measure λ will be assumed to dominate each of the Q_n's as well as the family $\{P_{\underline{\theta}}\}$.

Write $L_n(\epsilon)$ for the metric entropy with inclusion obtained by using the class $\mathcal{C}$ and the measure Q_n. For ease of notation write P for $P_{\underline{\theta}_0}$. Suppose that there exists $\delta \in D_0(\mathcal{C}, \lambda)$ such that

$$\sup_{C \in \mathcal{C}} |\sqrt{n}\,[Q_n(C) - P(C)] - \delta(C)| \to 0. \tag{*}$$

Then given that F_n has the appropriate measurability properties it can be shown, by slightly modifying the proof of *Dudley* [1978], that a sufficient condition for $\sqrt{n}\,[F_n - Q_n] \xrightarrow{\mathcal{D}} G_P$ as random elements of $D_0(\mathcal{C}, \lambda)$ is

$$\lim_{\alpha \to 0} \limsup_{n \to \infty} \int_0^\alpha L_n(\epsilon^2)^{1/2} d\epsilon = 0. \tag{**}$$

For example, the argument given in Section 3, for the case where $\mathcal{C}$ consists of all intervals $(-\infty, x]$ in $\mathbf{R}$, demonstrates that the uniform convergence in (**) holds for such a $\mathcal{C}$. The same argument extends to the multivariate case. We therefore have a result which will apply to the classical multidimensional empirical distribution functions.

Once we have that $\sqrt{n}\,[F_n - Q_n] \xrightarrow{\mathcal{D}} G_P$, it follows immediatley from (*) that $\sqrt{n}\,[F_n - P] \xrightarrow{\mathcal{D}} \delta + G_P$; that is, a drift term has been added to the limit Gaussian process. This then appears in the limit distribution for the test statistic under the alternatives $\{Q_n\}$; it is only necessary to replace G_P by $\delta + G_P$ to evaluate the limiting

power under alternatives of this type.

The added drift term in the limit also appears in the work of *Neuhaus* [1973, 1976a, 1976b] who considered sequences of contiguous alternatives. Some conditions under which contiguous alternatives may be expressed in the form (*) can be deduced from the works just cited.

In the one dimensional case, convergence of $\sqrt{n}\,(F_n - Q_n)$ can also be proved by more well-known methods. Working with the space $D[0,1]$, *Chibisov* [1965] obtained the result by means of the usual representation of an empirical distribution function in terms of that for uniformly distributed random variables. His method even incorporates the use of different metrics on $D[0,1]$, metrics which are more sensitive to the behaviour at 0 and 1.

Finally, to come back to the situation with which we began, notice that a sufficient condition for (*) can be expressed in terms of the densities $\xi_n^2 := (dQ_n/d\lambda)$. Exactly the same argument as in Example 2.3 would show that if $\sqrt{n}\,[\xi_n - \xi(\theta_0)]$ converges in $L^2(\lambda)$ norm to Δ then (*) holds with $\delta(C) := 2\int_C \xi(\theta_0)\,\Delta d\lambda$. In particular, if the underlying model can be embedded in a two parameter family $P(\cdot\,; \underline{\theta}, \underline{\eta})$, with $P_{\underline{\theta}}(\cdot) = P(\cdot\,; \underline{\theta}, \underline{\eta}_0)$, and the densities for this family satisfy the quadratic differentiability condition (as a function two variables $\underline{\theta}$ and $\underline{\eta}$ at $(\underline{\theta}_0, \underline{\eta}_0)$ then (*) would be obtained by considering $Q_n(\cdot) := P(\cdot\,; \underline{\theta}_0, \underline{\eta}_0 + n^{-1/2}\underline{\gamma})$ for some fixed $\underline{\gamma}$. Such sequences of alternatives were built into the model considered by *Durbin* [1973]; the end effect there was also to add a drift term onto the limit Gaussian process.

7. Minimum Distance Estimators

In general the procedures for testing a hypothesis and estimating parameters specified by the hypothesis often represent complementary aspects of a statistical model. This is indeed the case for the model considered in this paper: goodness-of-fit can be tested by the magnitude of a minimized distance, and the parameter can be estimated by the value of $\underline{\theta}$ at which the minimum is achieved (or at least where some value suitably close to the infimum is attained). In the papers of *Blackman* [1955] and *Bolthausen* [1977], as well as in the basic work of *Wolfowitz* (culminating in his 1957 paper), the properties of such minimum distance estimators have been investigated for a number of cases. In this section a method for obtaining the limiting distribution for the minimizing value of $\underline{\theta}$ in the situation of Theorem 4.2 will be described; the corresponding results for the cases of norms depending on the parameter may be derived analogously.

The simplest case occurs when the function $\underline{t} \to \| G - \langle \underline{t}, \underline{D} \rangle \|$ achieves its minimum at a unique value of $\underline{t}$, for almost all sample paths of G. The functional $\mu(\cdot)$ which associates with $x \in X$ a value of $\underline{t}$ minimizing $\| x - \langle \underline{t}, \underline{D} \rangle \|$ will then be G almost surely continuous; the argument can be based on the convexity of $\underline{t} \to \| x - \langle \underline{t}, \underline{D} \rangle \|$ cf. Proposition 3.2 of *Bolthausen* [1977]. If then $\underline{\theta}_n^*$ is a measurable function for which $\| F_n - F(\underline{\theta}_n^*) \| = \inf_{\underline{\theta}} \| F_n - F(\underline{\theta}) \|$ (an extra $o_p(1/\sqrt{n})$ term could be added to the

infimum without changing the result), the same argument as for the proof of Theorem 4.2 can be used to prove that $\sqrt{n}\,(\underline{\theta}_n^* - \underline{\theta}_0) \xrightarrow{\mathcal{D}} \mu(G)$; in Step III the functional $x \to \mu(x)$ rather than the infimum functional should be applied. [The last part of the argument to Step II would also need some modification but, since a more general result will be described below, the details will be omitted.]

A slight complication occurs if the infimum of $\| G - \langle \underline{t}, \underline{D} \rangle \|$ need not be achieved (almost surely) at a unique point. The most that could be asserted in that case would be that $\sqrt{n}\,(\underline{\theta}_n^* - \underline{\theta}_0)$ should behave asymptotically like one of these minimizing values; but this falls short of a true limit theorem for $\sqrt{n}\,(\underline{\theta}_n^* - \underline{\theta}_0)$. To overcome this difficulty I propose considering the entire set of minimizing values, and proving a limit theorem for this random set.

Define $M_n := \{\underline{\theta} \in \Theta: \| F_n - F(\underline{\theta}) \| \leqslant \inf_{\underline{\theta}'} \| F_n - F(\underline{\theta}') \| + \eta_n/\sqrt{n}\}$ where $\{\eta_n\}$ denotes some fixed sequence of random variables, with $\eta_n = o_p(1)$, chosen to ensure that M_n be non-empty. [This avoids the problem of the infimum not being achieved.] The limit result will assert the existence of a sequence of random compact convex sets $\{K_n\}$ which converge in distribution to the minimum set of $\underline{t} \to \| G - \langle \underline{t}, \underline{D} \rangle \|$, and for which $M_n \subseteq \underline{\theta}_0 + n^{-1/2} K_n$ with inner probability converging to one. The use of inner probability seems necessary since there is no guarantee that the set $\{M_n \subseteq \underline{\theta}_0 + n^{-1/2} K_n\}$ be measurable. The precise description of the random sets $\{K_n\}$ requires some preliminary definitions.

The class of all compact, convex, non-empty subsets of $\mathbf{R}^s$ will be denoted by $\mathcal{K}$. For each $x \in \mathcal{X}$ and each $\beta \geqslant 0$ define

$$f(x, \underline{t}) := \| x - \langle \underline{t}, \underline{D} \rangle \|,$$
$$m(x) := \inf_{\underline{t}} f(x, \underline{t}),$$

and

$$K(x, \beta) := \{\underline{t} \in \mathbf{R}^s : f(x, \underline{t}) \leqslant m(x) + \beta\}.$$

First observe that both $f(\cdot, \underline{t})$ and $m(\cdot)$ are $\mathcal{B}_0$ measurable – the infimum may clearly be replaced by an infimum over a countable dense set of $\underline{t}$ values. When the derivative $\underline{D}$ is non-singular, the function $\underline{t} \to f(x, \underline{t})$ becomes unbounded as $|\underline{t}| \to \infty$. It follows that $K(\cdot, \beta)$ defines a map from $\mathcal{X}$ into $\mathcal{K}$.

The natural topology on $\mathcal{K}$ is that generated by the Hausdorff metric defined by

$$d(K_1, K_2) := \inf \{\delta > 0: K_1^\delta \supseteq K_2 \text{ and } K_2^\delta \supseteq K_1\}$$

where K^δ denotes the closed set of points at distance less than or equal to δ from K. A general reference for this topology is *Eggleston's* [1977] book (note that he has used a slightly different metric).

7.1 Lemma

For each fixed $\beta \geqslant 0$, then map $x \to K(x, \beta)$ is $\mathcal{B}_0$-Borel measurable.

Proof: Notice that as $\beta \downarrow 0$ the sets $K(x, \beta)$ decrease to $K(x, 0)$, which implies convergence with respect to the Hausdorff metric topology. Thus it suffices to consider only the case where $\beta > 0$. This case is simpler because each $K(x, \beta)$ then has nonempty interior.

Since the Hausdorff metric induces a separable topology on $\mathcal{K}$, it suffices to show that the inverse image of each closed ball in $\mathcal{K}$ belongs to $\mathcal{B}_0$. Now the inverse image of the closed ball with centre K_0 and radius $r > 0$ may be represented as the intersection of the two sets $A_1 := \{x \in \mathcal{X} : K(x, \beta) \subseteq K_0^r\}$ and $A_2 := \{x \in \mathcal{X} : K_0 \subseteq K(x, \beta)^r\}$. Consider A_1 first.

Let T_0 be a countable dense subset of the complement of K_0^r. I assert that $A_1 = \bigcap_{\underline{t} \in T_0} \{x \in \mathcal{X} : f(x, \underline{t}) > m(x) + \beta\}$. That the latter set contains A_1 is clear. On the other hand suppose that $x \notin A_1$. Then K_0^r does not contain $\operatorname{int} K(x, \beta)$, for otherwise it would also contain $K(x, \beta)$ which is the closure of its interior. The nonempty open set $\operatorname{int} K(x, \beta) \setminus K_0^r$ thus contains a point of T_0: that is, $f(x, \underline{t}) \leqslant m(x) + \beta$ for at least one $t \in T_0$.

For the set A_2 start with a countable dense subset $\{\underline{t}_1, \underline{t}_2, \ldots\}$ of K_0 and choose T_n to be a countable dense subset of the closed ball with centre $\underline{t}_n$ and radius r. Simple continuity arguments show that

$$A_2 = \bigcap_{n=1}^{\infty} \bigcap_{p=1}^{\infty} \bigcup_{\underline{t} \in T_n} \{x \in \mathcal{X} : f(x, \underline{t}) \leqslant m(x) + \beta + p^{-1}\}.$$

The $\mathcal{B}_0$ measurability of the functions $f(\cdot, \underline{t})$ and $m(\cdot)$, together with the countable nature of the unions and intersections involved in the above representations for A_1 and A_2, leads us to the desired conclusion. □

This takes care of any problems regarding measurability which might have arisen in the proof of the main result of this section.

7.2 Theorem

Under the conditions of Theorem 4.2 there exists a sequence of real number $\beta_n \downarrow 0$ satisfying

(i) $\mathbf{P}_*\{M_n \subseteq \underline{\theta}_0 + n^{-1/2} K(G_n, \beta_n)\} \to 1$,

(ii) $K(G_n, \beta_n) \xrightarrow{\mathcal{D}} K(G, 0)$ as random elements of $\mathcal{K}$ under its Hausdorff metric topology.

Proof: Most of the proof is already contained in Lemma 4.1 and Steps I and II of Theorem 4.2; another variant of the continuous mapping theorem will be needed to replace Step III and complete the argument. Once again it will suffice to prove a stronger result (this time convergence in probability) for a version of the process satisfying $\bar{G}_n \xrightarrow{\text{a.s.}} \bar{G}$. With this in mind, let us ease the notation and omit the bars denoting the

use of a different version. I take up the proof from midway through Step I of Theorem 4.2.

From the inequality

$$\| F_n - F(\underline{\theta}) \| \geqslant \frac{1}{2} C \, | \, \underline{\theta} - \underline{\theta}_0 \, | - \| F_n - F(\underline{\theta}_0) \|,$$

valid for $\underline{\theta} \in N_1$, it follows that (with inner probability tending to one)

$$M_n \subseteq \underline{\theta}_0 + n^{-1/2} L_n,$$

where $L_n := \{\underline{t} \in \mathbf{R}^s : | \, \underline{t} \, | \leqslant (4 \| G_n \| + 2\eta_n)/C\}$. Define

$\Gamma_n := \sup_{\underline{t} \in L_n} | \sqrt{n} \| F_n - F(\underline{\theta}_0 + \underline{t}/\sqrt{n}) \| - \| G_n - \langle \underline{t}, \underline{D} \rangle \| |$. By the argument in Step II this quantity is of order $o_p(1)$. We can therefore find real numbers $\gamma_n \downarrow 0$ for which $\mathbf{P}\{\Gamma_n > \gamma_n\} \to 0$.

Also, since $\eta_n = o_p(1)$, there exist constants $\delta_n \downarrow 0$ satisfying

$$\mathbf{P}\{\eta_n > \delta_n\} \to 0.$$

Finally choose constants $\epsilon_n \downarrow 0$ for which

$$\mathbf{P}\{\| G_n - G \| > \epsilon_n\} \to 0.$$

Define $\beta_n := \max\{2\gamma_n + \delta_n, 2\epsilon_n\}$. To complete the proof we have only to show that:

a) $K(G_n, \beta_n) \xrightarrow{\mathbf{P}} K(G, 0)$;

b) if $\underline{\tau} \in L_n$ and $\underline{\theta}_0 + n^{-1/2}\underline{\tau} \in M_n$ and $\Gamma_n \leqslant \gamma_n$ and $\eta_n \leqslant \delta_n$ then $\underline{\tau} \in K(G_n, \beta_n)$.

For the proof of a) first notice that $| f(x, \underline{t}) - f(y, \underline{t}) | \leqslant \| x - y \|$ for all $x, y \in \mathcal{X}$ and $\underline{t} \in \mathbf{R}^s$, and hence $| m(x) - m(y) | \leqslant \| x - y \|$. Thus when $\| G_n - G \| \leqslant \epsilon_n$ we have

$$\begin{aligned} K(G, 0) &= \{\underline{t} : f(G, \underline{t}) \leqslant m(G)\} \\ &\subseteq \{\underline{t} : f(G_n, \underline{t}) \leqslant m(G_n) + 2 \| G_n - G \|\} \\ &\subseteq K(G_n, \beta_n). \end{aligned}$$

On the other hand, for any given $\alpha > 0$ and sample point ω there exists a value $\lambda_0 = \lambda_0(\omega)$ for which $K(G(\omega), \lambda) \subseteq K(G(\omega), 0)^\alpha$ whenever $\lambda \leqslant \lambda_0$ [if a decreasing sequence $\{K_n\}$ of compact sets has intersection contained in a given open set U then $K_n \subseteq U$ eventually]. Thus if $\{\lambda_n\}$ forms a sequence of random variables converging a.s. to zero it follows (after a measurability argument similar to that for Lemma 7.1) that

$$\mathbf{P}\{K(G, \lambda_n) \subseteq K(G, 0)^\alpha\} \to 1.$$

Apply this with $\lambda_n = \beta_n + 2 \| G_n - G \|$ to see that

$$\begin{aligned} K(G_n, \beta_n) &= \{\underline{t}: f(G_n, \underline{t}) \leqslant m(G_n) + \beta_n\} \\ &\subseteq \{\underline{t}: f(G, t) \leqslant m(G) + \beta_n + 2 \| G_n - G \|\} \\ &\subseteq K(G, 0)^{\alpha} \text{ with probability tending to one.} \end{aligned}$$

Part a) is thus proven.

For part b) start from the inequality for $\underline{t} \notin L_n$

$$\begin{aligned} \| G_n - \langle \underline{t}, \underline{D} \rangle \| &\geqslant C \mid \underline{t} \mid - \| G_n \| \\ &> 3 \| G_n \| \\ &\geqslant \| G_n - \langle \underline{0}, \underline{D} \rangle \| \end{aligned}$$

to deduce that

$$m(G_n) = \inf_{\underline{t} \in L_n} \| G_n - \langle \underline{t}, \underline{D} \rangle \|.$$

Then, since $\Gamma_n \leqslant \gamma_n$ and $\eta_n \leqslant \delta_n$,

$$\begin{aligned} m(G_n) &\geqslant \inf_{\underline{t} \in L_n} \sqrt{n} \| F_n - F(\underline{\theta}_0 + \underline{t}/\sqrt{n}) \| - \gamma_n \\ &\geqslant \sqrt{n} \| F_n - F(\underline{\theta}_0 + \underline{\tau}/\sqrt{n}) \| - \delta_n - \gamma_n \\ &\geqslant \| G_n - \langle \underline{\tau}, \underline{D} \rangle \| - \gamma_n - \delta_n - \gamma_n, \end{aligned}$$

implying that $\underline{\tau} \in K(G_n, \beta_n)$. □

The procedure used to define the constants β_n has the curious feature that it depends on the choice of a particular version of the $\{G_n\}$; the rate of convergence in probability specified by ϵ_n need not be the same for all versions. The final assertion of the theorem, however, must be valid for all versions. This apparent paradox may be partly explained by results of *Strassen* [1965] and *Dudley* [1968] which connect the ϵ_n's with the rate of convergence to zero of the Prohorov distance between $\mathbf{P}G_n^{-1}$ and $\mathbf{P}G^{-1}$; this rate must indeed be the same for all versions.

The result in the case where $\| G - \langle \underline{t}, \underline{D} \rangle \|$ achieves its minimum (almost surely) in a unique point can be recovered from Theorem 7.2 by noticing that, if $K_n \in K$ and $K_n \to \{\underline{t}\}$ in the sense of the Hausdorff metric, then $\underline{t}_n \to \underline{t}$ for each sequence of points $\underline{t}_n \in K_n$. As *Bolthausen* [1977] has observed however, only in the case of the L^2 type norms is it at all easy to verify the condition of uniqueness of minimizing values.

8. The Method Based on Direct Estimation of the Parameter

Let us consider once more the other approach to goodness-of-fit testing where the estimation of $\underline{\theta}$ proceeds by a so-called efficient method. As an example start with a modified version of the situation treated by *Durbin* [1973, 1976].

The empirical distribution function F_n is based on n independent observations $X_1, \ldots, X_n$ on the distribution $F(\cdot, \underline{\theta}_0)$. As the space $\mathcal{X}$ use $D[-\infty, \infty]$ under its sup norm. The test statistic is derived from the random function $\sqrt{n}\,(F_n(\cdot) - F(\cdot, \hat{\underline{\theta}}_n))$ where $\hat{\underline{\theta}}_n$ is an estimator having the following form:

$$\sqrt{n}\,(\hat{\underline{\theta}}_n - \underline{\theta}_0) = n^{-1/2} \sum_{i=1}^{n} \underline{\ell}\,(X_i) + o_p(1). \qquad (*)$$

When $\mathbf{E}\,\underline{\ell}\,(X_i) = \underline{0}$ and the variance matrix $\operatorname{var}\underline{\ell}\,(X_i)$ exists the central limit theorem ensures asymptotic normality of $\hat{\underline{t}}_n := \sqrt{n}\,(\hat{\underline{\theta}}_n - \underline{\theta}_0)$, and consequently $\hat{\underline{t}}_n = O_p(1)$. Under the conditions that $G_n(\cdot) := \sqrt{n}[F_n(\cdot) - F(\cdot, \underline{\theta}_0)]$ converges in distribution and that $\underline{\theta} \to F(\cdot, \underline{\theta})$ is norm differentiable at $\underline{\theta}_0$, the same arguments as for Theorem 4.2 show that

$$\| \sqrt{n}\,(F_n(\cdot) - F(\cdot, \hat{\underline{\theta}}_n)) - (G_n(\cdot) - \langle \hat{\underline{t}}_n, \underline{D}(\cdot)\rangle) \| = o_p(1).$$

This time it will be the joint distribution of G_n and $\hat{\underline{t}}_n$ which determines the asymptotic behaviour of $G_n - \langle \hat{\underline{t}}_n, \underline{D}\rangle$. *Durbin* proved convergence in distribution of this random function by the standard uniform tightness plus convergence of fidis argument. As *Neuhaus* [1976a] later noted, the uniform tightness follows easily from the fact that both $G_n(\cdot)$ and $\langle \hat{\underline{t}}_n, \underline{D}(\cdot)\rangle$ satisfy the usual condition restraining the oscillation of their sample paths. The fidi convergence results from an application of the multidimensional form of the central limit theorem by writing $G_n(\cdot) = n^{-1/2} \sum_{i=1}^{n} d(X_i, \cdot)$, where $d(X_i, \cdot) = 1_{(-\infty, \cdot]}(X_i) - F(\cdot, \underline{\theta}_0)$. For $G_n(\cdot) - \langle \hat{\underline{t}}_n, \underline{D}(\cdot)\rangle$ then has the asymptotic form of a normed sum of i.i.d. random functions: $n^{-1/2} \sum_{i=1}^{n} [d(X_i, \cdot) - \langle \underline{\ell}\,(X_i), \underline{D}(\cdot)\rangle] + o_p(1)$.

Neuhaus [1973] used a similar representation together with a central limit theorem for sums of i.i.d. Hilbert space valued random elements to prove the corresponding weak convergence result for the Cramér-von-Mises type statistics.

It is thus apparent that much the same apparatus can be used for both approaches to constructing goodness-of-fit tests. Only in the last steps of the method do significant differences occur; whereas the minimum distance method essentially depends on variants of the continuous mapping theorem, the method with direct parameter estimation requires an asymptotic form for $\hat{\underline{\theta}}_n$ as in (*) together with a weak form of central limit theorem for sums of i.i.d. $\mathcal{X}$ valued random elements.

The differentiability conditions required for the second method can usually be interpreted as a means for obtaining differentiability in norm. Also hidden in this method

though are extra (differentiability) requirements needed to justify the form (*) for $\hat{\underline{\theta}}_n$; usually reference is made to the classical conditions of *Cramér* [1964] for asymptotic efficiency of maximum likelihood estimators (see for example the further comments of *Durbin* [1973] or *Csörgő/Burke* [1976]). In this respect it is interesting to note that *Le Cam* [1970] introduced the quadratic differentiability condition of Example 2.3 in his study of conditions needed to prove asymptotic normality of maximum likelihood estimators.

Perhaps the most interesting problem for either of these two methods remains that of freeing the limit distribution of the test statistic from dependence on unknown parameters. Apart from the procedure mentioned at the end of Example 4.4, the half-sample device of *Durbin* [1976] seems the most promising method to date, although it has been criticised on the grounds that it requires post-sampling randomisation. Maybe the corresponding problem in the area of the χ^2 goodness-of-fit tests [cf. *Chernoff/Lehmann; Watson*] could give some clue to a more satisfactory solution.

Acknowledgements

My thanks are due to Peter Gänßler who not only introduced me to the work of *Bolthausen* and *Neuhaus*, but also made valuable comments on the penultimate draft of this manuscript. I also benefitted from conversations with Winfried Stute and Georg Neuhaus.

References

Billingsley, P.: Convergence of Probability Measures. New York 1968.

Birch, M.W.: A new proof of the Fisher-Pearson theorem. Ann. Math. Statist. **35**, 1964, 817–824.

Blackman, J.: On the approximation of a distribution function by an empirical distribution. Ann. Math. Statist. **26**, 1955, 256–267.

Bolthausen, E.: Convergence in distribution of minimum distance estimators. Metrika **24**, 1977, 215–227.

Chernoff, H.: On the distribution of the likelihood ratio. Ann. Math. Statist. **25**, 1954, 573–578.

Chernoff, H., and *E.L. Lehmann*: The use of maximum likelihood estimates in χ^2 tests for goodness-of-fit. Ann. Math. Statist. **25**, 1954, 579–586.

Chibisov, D.M.: An investigation of the asymptotic power of the tests of fit. Theor. Prob. Appl. **10**, 1965, 421–437.

Cramér, H.: Mathematical Methods of Statistics. Princeton 1964.

Csörgő, M., and *M.D. Burke*: Weak approximations of the empirical process when parameters are estimated. Lect. Notes in Math. **566**, 1976, 1–16.

Csörgő, M., J. Komlós, P. Major, P. Révész and *G. Tusnády*: On the empirical process when parameters are estimated. Trans. Seventh Prague Conference 1974. Prague 1977, 87–97.

Darling, D.A.: The Cramér/Smirnov test in the parametric case. Ann. Math. Statist. **26**, 1955, 1–20.

Dudley, R.M.: Weak convergence of probabilities on non-separable metric spaces and empirical measures on Euclidean spaces. Ill. J. Math. **10**, 1966, 109–126.

–: Measures on non-separable metric spaces. Ill. J. Math. **11**, 1967, 449–453.

–: Distances of probability measures and random variables. Ann. Math. Statist. **39**, 1968, 1563–1572.

–: Sample functions of the Gaussian process. Ann. Probability **1**, 1973, 66–103.
–: Central limit theorems for empirical measures. Ann. Probability, **6**, 1978, 899–929.
Durbin, J.: Weak convergence of the sample distribution function when parameters are estimated. Ann. Statistics **1**, 1973, 279–290.
–: Kolmogorov-Smirnov tests when parameters are estimated. Lect. Notes in Math. **566**, 1976, 33–44.
Eggleston, H.G.: Convexity. Cambridge 1977.
Elker, J.D., D. Pollard and *W. Stute*: Glivenko-Cantelli theorems for classes of convex sets. Adv. Appl. Prob., 1979, (to appear).
Kac, M., J. Kiefer and *J. Wolfowitz*: On tests of normality and other tests of goodness of fit based on distance methods. Ann. Math. Statist. **26**, 1955, 189–211.
Le Cam, L.: On the assumptions used to prove asymptotic normality of maximum likelihood estimates. Ann. Math. Statist. **41**, 1970, 802–828.
Neuhaus, G.: Asymptotic properties of the Cramér-von Mises statistic when parameters are estimated. Proc. Prague Symp. on Asymp. Stat. Ed by Hájek. 1973, 257–297.
–: Weak convergence under continguous alternatives of the empirical process when parameters are estimated: the D_k approach. Lect. Notes in Math. **566**, 1976a, 68–82.
–: Asymptotic power properties of the Cramér-von Mises test under contiguous alternatives. J. Multiv. Analysis **6**, 1976b, 95–110.
–: Asymptotic theory of goodness of fit tests when parameters are present: a survey. Lecture at tenth European mtg of Statisticians. Leuven 1977.
Pollard, D.: Weak convergence on non-separable metric spaces. J. Austral. Math. Soc. (Series A) **28**, 1979a, 197–204.
–: General chi-square goodness-of-fit tests with data-dependent cells. Z. Wahrscheinlichkeitstheorie verw. Geb, 1979b, (to appear).
Pyke, R.: Applications of almost surely convergent constructions of weakly convergent processes. Lect. Notes in Math. **89**, 1969, 187–200.
–: Asymptotic results for rank statistics. Nonparametric Techniques in Statistical Inference. Ed. by Puri. Cambridge 1970, 21–37.
Rockafellar, R.T.: Convex Analysis. Princeton 1972.
Strassen, V.: The existence of probability measures with given marginals. Ann. Math. Statist. **36**, 1965, 423–439.
Watson, G.S.: Some recent results in chi-square goodness-of-fit tests. Biometrics **15**, 1959, 440–468.
Wichura, M.J.: On the construction of almost uniformly convergent random variables with given weakly convergent image laws. Ann. Math. Statist. **41**, 1970, 284–291.
Witting, H., and *G. Nölle*: Angewandte Mathematische Statistik. Stuttgart 1970.
Wolfowitz, J.: The minimum distance method. Ann. Math. Statist. **28**, 1957, 75–88.

Received July 6, 1978

Note added in proof:

Prof. Lucien Le Cam has kindly shown me some old lecture notes of his, dating from the late fifties, in which he proved results similar to some of those in my paper. He has also derived some of the properties of minimum distance estimators on pp. 103–107 of his "Théorie Asymptotique de la Décision Statistique" (University of Montreal Press, 1969).