Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

R.D. Gill

Non- and semi-parametric maximum likelihood estimators and the von Mis~~i~~es method (part I)

Department of Mathematical Statistics   Report MS-R8709   July

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

Non- and Semi-Parametric Maximum Likelihood Estimators and the von Mises Method (Part I)

Richard D. Gill
Centrum voor Wiskunde en Informatica,
Kruislaan 413, 1098 SJ Amsterdam

Revised Version, May 1987

After introducing the approach to von Mises derivatives based on compact differentiation due to REEDS (1976), we show how non-parametric maximum likelihood estimators can often be defined by solving infinite dimensional score equations. Each component of the score equation corresponds to the derivative of the log likelihood for a one-dimensional parametric submodel. By means of examples we show that it usually is not possible to base consistency and asymptotic normality theorems on the implicit function theorem. However (in Part II) we show for a particular class of models, that once consistency (in a rather strong sense) has been established by other means, asymptotic normality and efficiency of the non-parametric maximum likelihood estimator can be established by the von Mises method.

Key Words and Phrases: non-parametric maximum likelihood, von Mises method, compact differentiation, Hadamard differentiation, asymptotically efficient estimation.
Mathematics classification:
Primary: 62G05, 62G20.
Secondary: 60B12, 60F17, 46A05.

PART I

1. INTRODUCTION

In a large number of practical situations one meets with the following phenomenon. Estimators are derived in a non- or semi-parametric problem by appealing to some generalization of the maximum likelihood principle. When centred and scaled by $\sqrt{n}$ these estimators turn out to be asymptotically Gaussian (about the true parameter value) with a covariance structure which is of analogous form to the inverse Fisher information matrix in a parametric model. In fact the estimators are asymptotically efficient in the sense of achieving the asymptotic bounds of BEGUN et al. (1983); see also WELLNER (1985) or BICKEL et al. (1987).

Our aim in these notes is to offer an explanation for these coincidences. Some particular cases in which they occur are the following: estimation of an unknown distribution function by the empirical distribution function (based on n independent and identically distributed observations); estimation of an unknown distribution function by the Kaplan-Meier or product-limit estimator based on n censored survival times; estimation of cumulative or integrated intensities (hazard rates) in Markov or semi-Markov processes by the Aalen-Nelson estimator (empirical cumulative hazard function) based on possibly censored observation of the process; estimation of regression coefficients and integrated base-line hazard in Cox's (1972) regression model by Cox's maximum partial likelihood estimator; estimation of an unknown distribution function in VARDI's (1985) selection bias models (see GILL &

Report MS-R8709
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

Wellner (1986)); and so on. Of course there are also well-known models where non-parametric maximum likelihood fails completely, and some very important models where the question is completely open.

The above examples should make it clear that by a non-parametric model we really mean: a model with an infinite-dimensional parameter space, for example a space of distribution functions or cumulative hazard functions. By a parametric model we mean a model with finite dimensional (Euclidean) parameter. A semi-parametric model has components of both types. In the examples there are even more parallels between non-parametric and parametric maximum likelihood estimation. For instance computation of the non-parametric maximum likelihood estimator (NPMLE) reduces to a formal parametric MLE problem, with as many parameters as observations. The inverse observed Fisher information for this pseudo-problem typically turns out to yield a consistent estimate of the covariance structure of the NPMLE in the original problem.

In attempting to explain these coincidences between non-parametric and parametric MLE we take a deliberately naieve approach. We shall only consider asymptotic results for situations with n independent and identically distributed observations, letting n tend to ∞ (the i.i.d. case). We only consider maximum likelihood estimators which are solutions of the likelihood equations (or score equations): derivative of log likelihood equals zero. For our large sample results we rely on the δ-method, i.e. on first order Taylor expansion. We do need to apply this method in an infinite-dimensional context, i.e. as the von Mises method. Here we make use of Reeds' (1976) elegant theory of von Mises expansions and von Mises-differentiation based on the so-called *compact* or *Hadamard* derivative. However within this approach we make the naieve choice of topology on the space of distribution functions: namely that based on the supremum norm. Finally we make as many regularity assumptions $-$ on existence of derivatives of various kinds, on the legitimacy of the interchange of differentiation and integration, etc. $-$ as are needed to make the proofs work.

Because of all these self-imposed restrictions it is not surprising that our final result is rather weak: we can only show (for a certain type of model, and under many regularity conditions) that *if* an NPMLE is consistent in a certain strong sense, *then* it is asymptotically Gaussian and in fact efficient: the limiting covariance structure can be interpreted as the inverse Fisher information, and no better limiting distribution is possible. Typical examples suggest that consistency has to be established by direct arguments specific for each particular case. However we do at least in general have a form of *Fisher consistency,* which makes proper consistency plausible.

By restricting the tools we use and concentrating on special cases we only obtain weak results. Clearly a more powerful and abstract approach is needed to get a mathematically attractive theory. However our approach is at least fairly accessible and it does show that a general theory is worth establishing. Also it really does give an explanation for the coincidences we described right at the start. The explanation can be summarized as follows: a sensibly defined non-parametric maximum likelihood estimator will also be the maximum likelihood estimate in any parametric submodel which happens to include or pass through the point given by the NPMLE. For smooth parametric submodels the NPMLE solves the likelihood equations. So even in non-parametric problems we can sometimes consider the NPMLE as a solution of the likelihood equations (score function equals zero) corresponding to every parametric submodel passing through it. In fact in many examples the NPMLE is uniquely determined by this property, even when attention is restricted to a (sufficiently large) subfamily of parametric submodels. Now, supposing the NPMLE to be consistent, we can hope to identify its limiting distribution by imitating the traditional proof of asymptotic normality of the MLE, which is based on a first order Taylor expansion of the score function. Key roles are played by the facts that, at the true parameter value, the score function has expectation zero while its variance equals minus the expectation of its derivative. All these properties have analogues in the infinite dimensional case, and indeed we can carry through (in Part II) an analogue of the traditional proof.

In a number of problems the actual definition of the NPMLE has been the subject of much discussion. In these problems we are given a model for continuously distributed observations which does not have a single obvious analogue for the discrete case, while in order to define the NPMLE a discrete extension of the model seems to be required. Different discrete extensions sometimes lead to different NPMLE's. Our results suggest that, as far as large sample properties are concerned, one can better try to extend *score functions* in as *smooth* a way as possible than to try to extend the whole model is some natural way.

2. Von Mises calculus and compact differentiability

2.1. Gateaux, Hadamard or Frechet ?

In these notes differentiation in infinite-dimensional spaces will turn up in various guises. We are going to consider estimates as functions of the empirical distribution of the data, and hence need in order to apply the δ-method, to differentiate with respect to distribution functions. This is the idea of the von Mises method. Also we need to take derivatives of log likelihood and of score functions with respect to the parameters of our models, themselves distribution functions or such like. In fact the NPMLE will be considered as an implicitly defined function of the empirical distribution function of the data, namely as the solution of the likelihood equation (itself depending on model parameter and on empirical distribution function). Finally the theory of asymptotic information inequalities for estimation in semi-parametric models is based on differentiation in certain $\mathcal{L}^2$-spaces of root densities.

Here we follow Reeds (1976) and Fernholz (1983) (unfortunately Reeds' work is not widely available) in using *compact* or *Hadamard* differentiability to get a really useful von Mises theory. Just as Reeds we introduce it in an abstract setting which allows comparison with the more familiar *Gateaux* and *Frechet* derivatives. Excellent surveys of the whole field are given by Averbukh & Smolyanov (1967, 1968). Especially appropriate is the quotation from Tolstoy which opens the first paper: *"How simple and clear this is" thought Pierre, "How could I not have known this before"*.

Nice applications of Reeds' approach, in proving asymptotic results for the jackknife and the bootstrap, are contained in Reeds (1978) and Bickel & Freedman (1981; Sections 3 and 8) respectively. We will further discuss the application of compact differentiability to the bootstrap in Section 2.4. Van der Vaart (1987; Chapter 4) gives results on efficiency of compactly differentiable functionals of efficient estimators.

The abstract differentiability theory we are going to discuss is usually formulated in the context of *locally convex topological vector spaces.* The theory of weak convergence, which we want to link up with differentiability, also exists for random elements of topological spaces, but it is not familiar and it is also more complicated unless the space is actually a *metric space.* We shall therefore consider both theories in what might be considered as the intersection of these two classes of spaces, namely in *normed vector spaces.* The main applications will concern $\mathbb{R}^k$ and $D(\overline{\mathbb{R}})$, the space of cadlag (right continuous with left-hand limits) real functions on the extended real interval $[-\infty,\infty]$. We shall endow $D(\overline{\mathbb{R}})$ with the supremum norm, i.e. for $x \in D(\overline{\mathbb{R}})$, $\|x\| = \sup_t|x(t)|$. (Under the usual Skorohod metric, $D(\overline{\mathbb{R}})$ is not even a topological vector space: addition is not a continuous operation!). To avoid the separability and measurability problems which usually motivate the choice of the Skorohod metric for $D[0,1]$, we shall work with Pollard (1984) and Gaenssler's (1984) weak convergence theory, which generalizes that of Billingsley (1968) by allowing the σ-algebra on the space, by reference to which *random elements* of that space are defined, to be different from (and generally smaller than) the Borel σ-algebra. We shall return to this topic later.

So let B_1 and B_2 denote two normed vector spaces. In specific applications these spaces will often be complete, i.e. will be Banach spaces, but they will often not be separable. Let $\phi:B_1 \rightarrow B_2$ be some function. How should we define differentiability of ϕ at some point x of B_1? Differentiability means that ϕ can be well-approximated by a *continuous, linear* map near x. But the term "well-approximated" can be interpreted in many ways. Let us first define the "remainder" of such an approximation, and then give a whole class of ways of saying that this remainder is small close to x.

Definition 1. For given ϕ, given x, and a given continuous linear function $d\phi(x):B_1 \rightarrow B_2$ we define the remainder of ϕ at $x+h$, $\mathrm{Rem}(x+h)$, by

$$\phi(x+h) = \phi(x)+\mathrm{d}\phi(x).h+\mathrm{Rem}(x+h) \tag{1}$$

Here h varies in B_1, though if ϕ is only defined in some neighbourhood of x then $\mathrm{Rem}(x+h)$ is only defined for h in some neighbourhood of zero. Of course $\mathrm{d}\phi(x).h=0$ when $h=0$, and $\mathrm{Rem}(x+0)=0$ too. We will say that ϕ is differentiable at x, with derivative $\mathrm{d}\phi(x)$ at that point, if $\mathrm{Rem}(x+h)$ is of smaller order than h as h tends to 0:

DEFINITION 2. Let $\mathbb{S}$ be a collection of subsets of B_1, let $t\in\mathbb{R}$.
Then ϕ is $\mathbb{S}$-differentiable at x with derivative $\mathrm{d}\phi(x)$ if $\forall\ S\in\mathbb{S}$

$$\frac{\mathrm{Rem}(x+th)}{t} \to 0 \ \text{ as } \ t\to 0 \ \text{ uniformly in } \ h\in S\,. \tag{2}$$

Different choices of $\mathbb{S}$ now correspond to requiring the linear approximation of ϕ to be more or less uniformly good as one moves towards x from different directions h. Three important and common choices are given in the next definition:

DEFINITION 3.
When $\mathbb{S}$ = all singletons of B_1, ϕ is called Gateaux or directionally differentiable.
When $\mathbb{S}$ = all compact subsets of B_1, ϕ is called Hadamard or compactly differentiable.
When $\mathbb{S}$ = all bounded subsets of B_1, ϕ is called Frechet or boundedly differentiable.

Clearly bounded differentiability (of ϕ at x) implies compact differentiability, and that implies directional differentiability. The derivative $\mathrm{d}\phi(x)$ remains the same. In applications one often determines the form of the derivative by computing the Gateaux derivative acting on h, $\mathrm{d}\phi(x).h$, for a collection of directions h which span B_1. This in turn comes down to computing the ordinary derivative (with respect to $t\in\mathbb{R}$) of the mapping $t\to\phi(x+th)$, at the point $t=0$.

When $B_1=\mathbb{R}$ (with the usual topology) all three definitions of differentiability are equivalent. In $\mathbb{R}^k$, $k>1$, Hadamard and Frechet differentiability are equivalent and strictly stronger than Gateaux differentiability. More generally the three are all different. Note also that in $\mathbb{R}^k$, $k\geqslant 1$, Hadamard and Frechet differentiability are equivalent to ordinary differentiability. The continuous linear map $\mathrm{d}\phi(x)$ can be identified with the vector of partial derivatives $\frac{\partial\phi}{\partial x_i}(x)$, $i=1,\ldots, k$; each an element of B_2.

Reeds' major point is that in statistical applications where B_1 contains empirical and underlying distribution functions and $\phi(F_n)$ is some statistical quantity of interest, Gateaux differentiability of ϕ at the underlying or true distribution function F is too weak to be of any use at all in theorem proving (it only supplies a heuristic tool for suggesting *what* theorem could be proved), while Frechet differentiability is so strong that hardly any interesting statistical functionals ϕ are differentiable at all. On the other hand Hadamard differentiability is exactly attuned to statistical applications and nicely separates analytical considerations about ϕ from probabilistic considerations about F_n. SERFLING (1980; chapter 6) only works with Gateaux and Frechet differentiability. To get a useful Gateaux theory, he generalizes the concept to a notion of 'Stochastic differentiability'. However this comes close to assuming what has to be proved, namely 'the remainder term is $o_p(1)$'; more importantly, analytic and stochastic considerations are not separated. In this way the theory gains in power but loses in elegance. POLLARD (1985) has another interesting concept of stochastic differentiability.

To motivate our further exploration of Hadamard differentiability, let us state a preliminary version of the 'δ-method' theorem. Here weak convergence is understood in the usual (BILLINGSLEY, 1968) sense. In it, X_n might play the role of an empirical distribution function, considered as a random element of B_1, and μ would then be the true distribution function:

THEOREM 1. (THE δ-METHOD, PRELIMINARY VERSION.) *Suppose $\phi:B_1\to B_2$ is Hadamard differentiable at $\mu\in B_1$ and measurable with respect to the Borel σ-algebras on B_1 and B_2. Suppose X_n is a sequence of* *random elements of B_1 such that $n^{\frac{1}{2}}(X_n-\mu)\xrightarrow{\mathfrak{D}}Z$ (in B_1) and such that $n^{\frac{1}{2}}(X_n-\mu)$ is tight. Then*

$$n^{\frac{1}{2}}(\phi(X_n)-\phi(\mu))\xrightarrow{\mathfrak{D}}\mathrm{d}\phi(\mu).Z \quad (\text{ in } B_2). \tag{3}$$

The theorem is also true when the sequence $n^{\frac{1}{2}}$ is replaced by a sequence of positive real constants $a_n\to\infty$ as $n\to\infty$. In separable, complete spaces weak convergence implies tightness, but this is not generally true ! The proof of the theorem is left as an exercise for the reader. Use the definition of compact differentiability, drawing the following correspondences:

$$t \longleftrightarrow 1/\sqrt{n} \qquad x \longleftrightarrow \mu$$

$$x+th \longleftrightarrow X_n \qquad h \longleftrightarrow \sqrt{n}(X_n-\mu)$$

$$\frac{\mathrm{Rem}(x+th)}{t} \longleftrightarrow \sqrt{n}(\phi(X_n)-\phi(\mu))-\mathrm{d}\phi(\mu).\sqrt{n}(X_n-\mu)$$

As previously mentioned, we shall later need a generalized version of this theorem. However it allows us to highlight an important point. In a typical statistical application we start with a statistical quantity T_n considered as a function of an empirical distribution function. Subject to their containing some representation of F_n and of $T_n=\phi(F_n)$ for possible realizations of an empirical distribution function F_n, the actual choice of the spaces B_1 and B_2, and especially of their topologies, is up to us. Also the definition of ϕ acting on elements of B_1 which are not empirical distribution functions is up to us. Making the topology on B_1 finer (more open sets, and thereby less compact sets) makes Hadamard differentiability and measurability of ϕ easier to verify, but makes weak convergence and tightness of $n^{\frac{1}{2}}(X_n-\mu)$ harder to verify. So a delicate trade-off can be made between establishing analytical properties of ϕ and probabilistic properties of X_n, leading perhaps to a different choice of topology for each different statistical functional one considers. Reeds is a master in these matters. We shall ignore these possibilities by making a naieve choice of topology (based on the supremum norm) in all the examples we look at. The many arbitrary choices involved here are avoided in VAN ZWET's (1984) projection approach: $\phi(F_n)$ is optimally linearly approximated in $\mathcal{L}^2$ sense. A disadvantage is that the approximation is generally different for each n, and hard to evaluate explicitly. Also it depends on the precise probabilistic behaviour of F_n.

2.2. Properties of Hadamard differentiation

Characterizations of differentiability.
Always taking $t\in\mathbb{R}$ and $h_n, h\in B_1$, we have two very useful equivalent definitions of Hadamard differentiability. These are that ϕ is Hadamard differentiable at x with derivative $\mathrm{d}\phi(x)$ if and only if

$$\frac{\mathrm{Rem}(x+th_n)}{t}\to 0 \quad \forall\ t\to 0,\ \forall\ h_n\to h\in B_1 \tag{4}$$

and if and only if

$$\frac{\mathrm{Rem}(x+th_n)}{t}\to 0 \quad \forall\ t\to 0,\ \forall \text{ compact } K\subseteq B_1 \text{ and sequences } h_n \text{ with } d(h_n,K)\to 0\,. \tag{5}$$

Here $d(h,K)=\inf_{k\in K}\|h-k\|$. One can also replace "$t$" by elements of a sequence t_n. Also one can restrict attention to $t_n>0$ in each case, taking limits as just $n\to\infty$. These characterizations are related to the fact that in line 2 of Definition 3, $\mathbb{S}$ can be replaced by the larger class of all *sequentially compact* subsets of B_1.

Differentiation tangentially to a subspace.
We shall find it extremely useful to consider a weaker kind of Hadamard differentiability in which we only consider, in (4), sequences $h_n \in B_1$ with limits $h \in H$ where H is a subspace of B_1. We say then that ϕ is Hadamard differentiable (at x) *tangentially to* the subspace H : taking again $t \in \mathbb{R}$ and $h_n \in B_1$, we require

$$t^{-1}\text{Rem}(x+th_n)\to 0, \ \forall \ t\to 0, \ \forall \ h_n\to h\in H . \tag{6}$$

This is stronger than supposing ϕ to be differentiable (at x) *inside* or *along* or restricted to h_n in the subspace H. We will also apply definition (6) in the case when ϕ is defined on some subset $E \subseteq B_1$ (generally not a subspace itself), but possessing a *tangent space* H at x: for all $h \in H$ there exist $h_n \to h$, $t_n(\in \mathbb{R}_+)\to 0$, such that $x+t_nh_n \in E \ \forall \ n$. When ϕ is differentiable tangentially to H, its derivative $\mathrm{d}\phi(x)$ is only defined as a continuous linear map from H to B_2. However when B_2 is a Euclidean space, extensions from B_1 to B_2 exist (Hahn-Banach theorem).

The chain rule.
A most important property of Hadamard differentiation is that it satisfies the chain rule: if $\phi{:}B_1\to B_2$ and $\psi{:}B_2\to B_3$ are Hadamard differentiable at $x \in B_1$ and $\phi(x)\in B_2$ respectively, then $\psi\circ\phi{:}B_1\to B_3$ is Hadamard differentiable at x with derivative $\mathrm{d}\psi(\phi(x)).\mathrm{d}\phi(x)$ (a continuous linear map from B_1 to B_3). In fact Hadamard differentiability is the weakest form of differentiation which satisfies the chain rule, and yet another equivalent definition is: ϕ is differentiable at x if and only if for all $\psi : \mathbb{R}\to B_1$ which are differentiable (in the ordinary sense) at 0 and satisfy $\psi(0)=x$, $\phi\circ\psi : \mathbb{R}\to B_2$ is also differentiable (in the ordinary sense) with derivative $\mathrm{d}\phi(x).\mathrm{d}\psi(0)$.

The chain rule also holds for Hadamard differentiation tangentially to a subspace provided the subspaces match up properly.

Inverse and implicit function theorems.
Since we are going to study estimators which are implicitly defined as solutions of an estimating equation, it is very natural to hope that an abstract version of the implicit function theorem will be applicable. Supposing $\psi{:}B_1\times B_2\to B_2$ to be a given function, the implicit function theorem gives conditions for *existence* and *differentiability* of a mapping $\phi{:}B_1\to B_2$ which supplies a solution $y\in B_2$ to the equation $\psi(x,y)=0$, for any given $x\in B_1$: so ϕ must satisfy $\psi(x,\phi(x))=0$ (perhaps just in the neighbourhood of a particular point $x_0\in B_1$). Such a theorem also identifies the derivative of ϕ in terms of the partial derivatives $\mathrm{d}_1\psi$ and $\mathrm{d}_2\psi$ of ψ with respect to x and y : One expects

$$\mathrm{d}\phi(x) = -[\mathrm{d}_2\psi(x, \phi(x))]^{-1}\mathrm{d}_1\psi(x, \phi(x)) .$$

REEDS (1976) gives a version of such a theorem for Hadamard differentiation. He notably requires B_2 to be a Banach space and ψ to be *continuously* differentiable (with respect to both arguments jointly) in a neighbourhood of (x_0,y_0) where $\psi(x_0,y_0)=0$. Continuous differentiability means that the derivative $\mathrm{d}\psi(x,y)$ varies continuously (with respect to the topology of uniform convergence on compact subsets of $B_1\times B_2$; see REEDS (1976) Appendix A) as the point (x,y) varies at which the derivative is taken. By means of some examples we later show that such a theorem will not be applicable to the NPMLE in the problems which motivated this study; at least not when the naieve choice of topology is made: continuous differentiability fails to hold. This is a very delicate matter; and there are errors concerning exactly this point in REID (1981) and in CROWLEY & TSAI (1985). REEDS (1976; Appendix A) shows that with *continuous* differentiability, Gateaux, Hadamard and Frechet theories more or less coincide. We did not succeed in getting around this problem by use of a more sophisticated topology. However Reeds makes impressive use of the implicit function theorem when studying (finite-dimensional) M-estimators. (He also makes some confusing errors. See HEESTERMAN (1987) for a corrected presentation).

An alternative and far less deep type of implicit function theorem is used by FERNHOLZ (1983). By explicitly assuming existence and a kind of pre-differentiability of the solution ϕ, she obtains differentiability and identifies the derivative as before under far weaker conditions on ψ. In particular ψ need only be differentiable at the point (x_0,y_0). We essentially take this approach (the other having failed), though since pre-differentiability is really as hard to verify as differentiability itself, we prefer for simplicity to assume that too!

Similar remarks to the above can be made on the subject of *inverse* function theorems, concerning the existence, differentiability, and identification of the derivative of an inverse $\phi=\psi^{-1}:B_1\rightarrow B_2$ of a given mapping $\psi : B_2\rightarrow B_1$.

The delta method.

Here we discuss the notion of weak convergence of Gaenssler (1984) and Pollard (1984), and finally present a definitive version of Theorem 1.

Let B be a normed vector space endowed with a σ-algebra $\mathfrak{B}$, such that

$$\mathfrak{B}'\subset\mathfrak{B}\subset\mathfrak{B}'',$$

where $\mathfrak{B}'$ and $\mathfrak{B}''$ are the σ-algebras generated by the *open balls* and the *open sets* respectively of B. Thus $\mathfrak{B}''$ is the Borel σ-algebra; when B is separable, $\mathfrak{B}' = \mathfrak{B}''$.

Definition 4. (weak convergence). Let X_n *be a sequence of random elements of* $(B,\mathfrak{B})$ *and let* X *be another random element of that space. We say* X_n *converges weakly (or in distribution) to* X, $X_n\xrightarrow{\mathfrak{D}}X$, *iff* $Ef(X_n)\rightarrow Ef(X)$ *for all bounded, continuous, measurable* $f:B\rightarrow\mathbb{R}$.

It turns out that all the usual weak convergence theorems remain valid under this broader definition, provided that X takes values in a separable subset of B; this applies in particular to the *continuous mapping theorem* and to the Skorohod - Dudley - Wichura *almost-sure representation theorem*, which we now state:

Theorem 2. (Skorohod-Dudley-Wichura). *Suppose* $X_n\xrightarrow{\mathfrak{D}}X$ *in* $(B,\mathfrak{B})$, *where* X *takes values in a separable subset of* B. *Then there exists a sequence* X_n', X' *defined on a single probability space, such that* $X_n'\overset{\mathfrak{D}}{=}X_n$ *for all* n, $X'\overset{\mathfrak{D}}{=}X$, *and* $X_n'\rightarrow X'$ *almost surely.*

Of course, $X'\overset{\mathfrak{D}}{=}X$ means that the probability measures they induce on $(B,\mathfrak{B})$ coincide; equivalently (this is a theorem!) $Ef(X')=Ef(X)$ for all bounded, continuous, measurable real f.

Considering Theorem 2 and the characterization (4) of differentiability by convergence of sequences enables us to prove the following theorem, in which both the δ-method (9) and a 'weak Bahadur representation' (8) are given:

Theorem 3. (δ-method, final version). *Suppose* $\phi:B_1\rightarrow B_2$ *is compactly differentiable at a point* $\mu\in B_1$ *and measurable with respect to* σ*-algebras* $\mathfrak{B}_1$ *and* $\mathfrak{B}_2$ *(each nested between the open-ball and Borel* σ*-algebras). Suppose* X_n *is a sequence of random elements of* B_1 *such that* $Z_n=n^{\frac{1}{2}}(X_n-\mu)\xrightarrow{\mathfrak{D}}Z$ *in* B_1, *where the distribution of* Z *is concentrated on a separable subset of* B_1. *Suppose addition:* $B_2\times B_2\rightarrow B_2$ *is measurable. Then*

$$(n^{\frac{1}{2}}(X_n-\mu),\ n^{\frac{1}{2}}(\phi(X_n)-\phi(\mu))-\mathrm{d}\phi(\mu)\cdot n^{\frac{1}{2}}(X_n-\mu))\ \xrightarrow{\mathfrak{D}}(Z,0) \tag{7}$$

in $B_1\times B_2$ *and consequently (in particular)*

$$n^{\frac{1}{2}}(\phi(X_n)-\phi(\mu))-\mathrm{d}\phi(\mu)\cdot n^{\frac{1}{2}}(X_n-\mu)\xrightarrow{P}0, \tag{8}$$

$$n^{\frac{1}{2}}(\phi(X_n)-\phi(\mu))\xrightarrow{\mathfrak{D}}\mathrm{d}\phi(\mu)\cdot Z \tag{9}$$

REMARK 1. Measurability of $\mathrm{d}\phi(\mu){:}B_1\rightarrow B_2$ can be shown to follow from measurability of ϕ. The theorem is also true when ϕ is only differentiable tangentially to a linear subspace on which the distribution of Z is concentrated, provided its derivative is continuous, linear and measurable on all of B_1.

REMARK 2. Continuing on measurability questions, we note that in the theorem we should have specified σ-algebras on $B_1\times B_2$ and $B_2\times B_2$, and also a norm on $B_1\times B_2$. For $x=(x_1,x_2)\in B_1\times B_2$, we define $\|x\|=\max(\|x_1\|,\|x_2\|)$. We give product spaces their product σ-algebras. If B_1 and B_2 are $D(\bar{\mathbb{R}})^p$ and/or $\mathbb{R}^q$ for some finite p and q, and $\mathfrak{B}_1$ and $\mathfrak{B}_2$ are the open-ball σ-algebras, then all product σ-algebras are also the open-ball σ-algebras with respect to the max norm. These facts follow from the characterization of the open ball σ-algebra on $D(\bar{\mathbb{R}})$ as the σ-algebra generated by the *coordinate mappings* or *projections* $x\rightarrow x(t)$.

PROOF. We sketch the main part of the proof of this theorem, leaving measurability questions to the interested reader. Since $Z_n=n^{\frac{1}{2}}(X_n-\mu)\xrightarrow{\mathfrak{D}}Z$ in $(B,\mathfrak{B})$ with Z's distribution concentrated on a separable subset, there exist Z_n', Z' on a single probability space with $Z_n'\overset{\mathfrak{D}}{=}Z_n$, $Z'\overset{\mathfrak{D}}{=}Z$, and $Z_n'\rightarrow Z'$ a.s. Now define $X_n'=\mu+n^{-\frac{1}{2}}Z_n'$; we have $X_n'\overset{\mathfrak{D}}{=}X_n$ for each n. By compact differentiability of ϕ, and using (4) with $t_n=n^{-\frac{1}{2}}$, $h_n=Z_n'$, we have $n^{\frac{1}{2}}(\phi(X_n')-\phi(\mu))\rightarrow\mathrm{d}\phi(\mu)\cdot Z'$a.s.. Also by continuity of $\mathrm{d}\phi(\mu)$, $\mathrm{d}\phi(\mu)\cdot Z_n'\rightarrow\mathrm{d}\phi(\mu)\cdot Z'$ a.s.. Thus

$$(n^{\frac{1}{2}}(X_n'-\mu),\ n^{\frac{1}{2}}(\phi(X_n')-\phi(\mu))-\mathrm{d}\phi(\mu)\cdot n^{\frac{1}{2}}(X_n'-\mu))\ \rightarrow\ (Z',0)\text{ a.s.} \tag{10}$$

Since convergence a.s. implies convergence in distribution, and the left hand sides of (7) and (10) have the same distribution for each n, we get (7). □

A useful lemma

In many applications the mapping ϕ is only a priori defined on certain members of B_1, e.g. elements of $D(\bar{\mathbb{R}})$ which are actually *distribution functions.* One could set about choosing a particular extension to all of B_1 such that the hypotheses of Theorem 3 are satisfied in each particular application. The following lemma shows that the choice of extension is irrelevant: one need only verify (4) for sequences which can occur in practice: a differentiable extension exists iff (4) holds when it is needed.

LEMMA 1. *Suppose* $x\in E\subset B_1$, $\phi{:}E\rightarrow B_2$, *and* $\bar{E}$ *is a neighbourhood of* x. *Suppose there exists a continuous linear map* $\mathrm{d}\phi(x){:}B_1\rightarrow B_2$ *such that for all* $t_n\rightarrow 0$ $(t_n\in\mathbb{R})$ *and* $h_n\rightarrow h\in B_1$ *such that* $x_n=x+t_nh_n\in E$ *for all n,*

$$t_n^{-1}(\phi(x+t_nh_n)-\phi(x))\rightarrow\mathrm{d}\phi(x).h\quad\text{as}\quad n\rightarrow\infty\ .$$

Then ϕ *can be extended to* B_1 *in such a way that it is differentiable at x, and any such extension has derivative* $\mathrm{d}\phi(x)$ *at x.*

PROOF. The existence of an extension, differentiable at x, is easily established by the choice $\phi(x+h)=\phi(x)+\mathrm{d}\phi(x).h$ for $x+h\notin E$. The fact that for given $0\neq h\in B_1$ and for arbitrary $\epsilon>0$ and an arbitrary neighbourhood of h one can find $t'\in\mathbb{R}$ with $0<|t'|<\epsilon$ and h' in the neighbourhood with $x+t'h'\in E$ shows that this extension is indeed differentiable with derivative $\mathrm{d}\phi(x)$. The same fact shows also that any differentiable extension has the same derivative. □

A similar result can be given for differentiability tangentially to a subspace. We leave the details to the reader. Of course the derivative is then not unique outside the subspace.

2.3. Examples.

The following simple examples illustrate the different kinds of problem which can arise when applying the previous theory, and in particular Theorem 3 and Lemma 1, to proving asymptotic normality of a statistical quantity, considered as a function of the empirical distribution function.

The examples concern independent and identically distributed observations on the real line. We make this restriction because in later application to non-parametric maximum likelihood estimation we work with *parameters* which are distribution functions or cumulative hazard functions on the real line. The observations may be multivariate.

The specific examples we consider here are the sample median or another sample quantile, and the two-sample Wilcoxon test. Thus if F_n and G_m are empirical distribution functions based on independent random samples of size n and m from distributions F and G on $\mathbb{R}$ respectively, we look at asymptotic normality of $\phi(F_n)=F_n^{-1}(p)$, $p\in(0,1)$ and of $\phi(F_n,G_m)=\int_{-\infty}^{\infty}F_n(x)dG_m(x)$. We want to obtain these results by using only the well-known weak convergence of $n^{\frac{1}{2}}(F_n-F)$ in $D(\overline{\mathbb{R}})$ (and similarly for G_m) and differentiability of the function ϕ in each case. The first example (also considered by Reeds, 1976) is purely illustrative; however the second is relevant to non-parametric maximum likelihood estimation since the functional $(F,G)\rightarrow\int FdG$ plays an important role in very many of the examples from survival analysis, Markov processes, etc.

First of all we restate Donker's theorem - concerning weak convergence of $n^{\frac{1}{2}}(F_n-F)$ - in the generalized set up of Section 2.2. Let $X_1,...,X_n$ be i.i.d. real random variables with distribution function F and empirical distribution function F_n. Both are elements of $D(\overline{\mathbb{R}})$, which we endow with the *supremum norm* and the *open-ball* σ-algebra $\mathfrak{B}$. Then

$$n^{\frac{1}{2}}(F_n-F)\xrightarrow{\mathfrak{D}}B^0\circ F \text{ in } (D(\overline{\mathbb{R}}),\mathfrak{B})$$

where B^0 is the Brownian bridge on [0,1] (a Gaussian process with continuous sample paths and $\mathrm{cov}(B^0(s),B^0(t))=s(1-t)$ for $0\leqslant s\leqslant t\leqslant 1$). The sample paths of $B^0\circ F$ lie in the separable subspace of $D(\overline{\mathbb{R}})$ consisting of functions continuous except at the points of discontinuity of F.

This version of Donker's theorem is *stronger* than the usual version, in which weak convergence holds in $D(\overline{\mathbb{R}})$ endowed with the Skorohod topology and its Borel σ-algebra. In general, if $Z_n\xrightarrow{\mathfrak{D}}Z$ in ($D(\overline{\mathbb{R}})$, Skorohod), *and* Z has continuous sample paths, then $Z_n\xrightarrow{\mathfrak{D}}Z$ in ($D(\overline{\mathbb{R}})$, supremum norm). So for applications to other objects than empirical distribution functions, it is usually easy to switch from classical to generalized weak convergence.

The following corollary (to Theorem 3 and Donker's theorem) shows what happens when $B_2=\mathbb{R}$, and makes the connection to the well-known influence curve of robust statistics. (See e.g. Hampel et al., 1986).

Corollary 1. *Suppose $\phi:(D(\overline{\mathbb{R}}),\|.\|)\rightarrow\mathbb{R}$ is Hadamard differentiable at F and suppose $\phi(F_n)$ is a random variable, where F_n is the empirical distribution function based on n independent and identically distributed observations $X_1,...,X_n$ from a distribution F on $\mathbb{R}$. Then*

$$n^{\frac{1}{2}}(\phi(F_n)-\phi(F))\xrightarrow{\mathfrak{D}}\mathrm{d}\phi(F).Z \tag{11}$$

where $Z=B^0\circ F$ and B^0 is a Brownian bridge on [0,1]. *In fact* $\mathrm{d}\phi(F).Z$ *is a normally distributed random variable with mean zero and (finite) variance that of*

$$\mathrm{d}\phi(F).(F_1-F) = \mathrm{IC}(\phi;F,X_1),$$

the influence curve of $\phi(F_n)$ evaluated at $x=X_1$:

$$\mathrm{IC}(\phi;F,x) = \lim_{t\to 0}\frac{\phi((1-t)F+t\,1_{[x,\,1]})-\phi(F)}{t}.$$

Also

$$n^{\frac{1}{2}}(\phi(F_n)-\phi(F))-n^{-\frac{1}{2}}\sum_{i=1}^{n}\mathrm{IC}(\phi;F,X_i)\overset{P}{\to}0.$$

In fact ϕ need only be Hadamard differentiable, tangentially to the subspace of $D(\overline{\mathbb{R}})$ consisting of functions whose points of discontinuity fall in the set of discontinuity points of F. This fact is vital to our applications.

Consider a statistic T_n which is a p'th quantile of the empirical distribution function F_n; i.e.

$$F_n(T_n-)\leqslant p\leqslant F_n(T_n);\quad 0<p<1. \tag{12}$$

This inequality does not generally uniquely define T_n as a function of F_n but that will not be important. We do suppose that T_n *is* a function of F_n (i.e. is a symmetric function of the n observations). So we are given $T_n=\phi(F_n)$ for some function ϕ on the set of distribution functions, and (12) holds. (We suppose ϕ is measurable, but ignore measurability questions from now on.). The next lemma, in combination with lemma 1 and the remark following it, shows that ϕ can be extended to all of $D(\overline{\mathbb{R}})$ in such a way that it is differentiable at F, *tangentially* to the subspace of functions continuous at the p'th quantile of F, provided F itself has a positive derivative (in the ordinary sense) at its p'th quantile (which is consequently uniquely defined). To make the notation lighter we write x for F and shift both p and the p'th quantile to the origin.

Lemma 2. *Let $x\in D(\overline{\mathbb{R}})$ be fixed and nondecreasing, and satisfy $x(0)=0$, x is differentiable at 0 with positive derivative $x'(0)$. Let h_n be a sequence of elements of $D(\overline{\mathbb{R}})$ and t_n a sequence of elements of $\mathbb{R}^+$ such that $h_n\xrightarrow{\|\cdot\|}h$, $t_n\to 0$, and h is continuous at 0. Define $x_n=x+t_nh_n$ and suppose $\theta_n\in\mathbb{R}$ satisfies*

$$x_n(\theta_n-)\leqslant 0\leqslant x_n(\theta_n)\ \ \forall_n. \tag{13}$$

Then

$$\psi_n = t_n^{-1}\theta_n\to -h(0)/x'(0) \text{ as } n\to\infty.$$

Before proving the Lemma, we illustrate the result by a sketch of the behaviour of x_n and x near the origin. Each coordinate axis has been rescaled by a factor $1/t_n$.

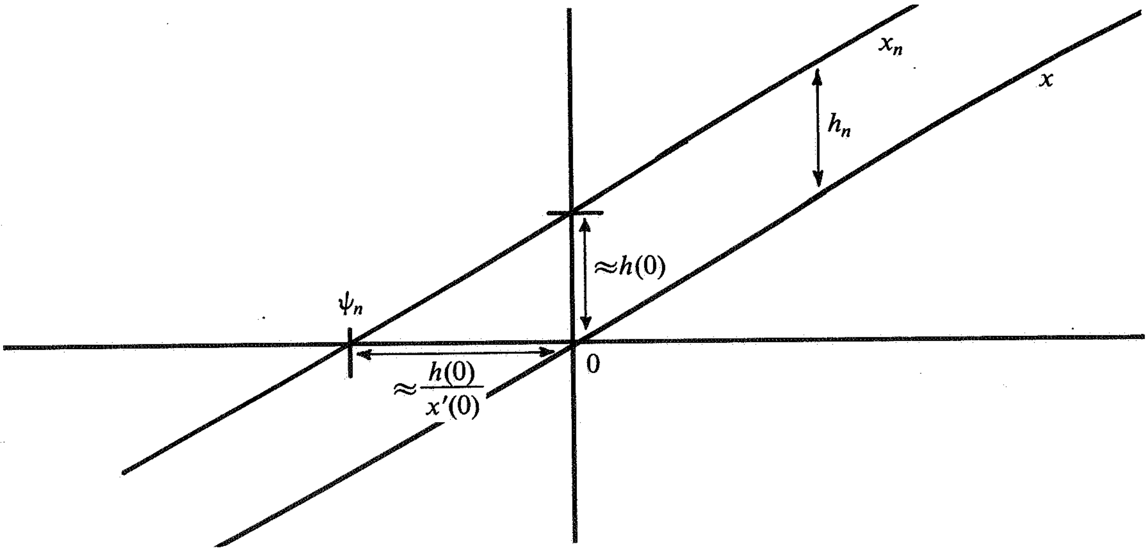

FIGURE 1. Derivative of $\phi(x)=x^{-1}(0)$

PROOF OF LEMMA 2. Suppose that we have already established that the rescaling in Figure 1 is legitimate; i.e. lim sup$|\psi_n|<M$ for some $M<\infty$. Consider $t_n^{-1}x_n(t_nu)$ for $u\in[-M,M]$. We have for $u\in[-M,M]$

$$x_n(t_nu) = x(t_nu)+t_nh_n(t_nu) \tag{14}$$

$$x(t_nu) = t_nu.x'(0)+o(t_n) \text{ uniformly in } u. \tag{15}$$

So substituting (15) in (14) and (14) in (13) with $u=\psi_n$, $h_n=t_n\psi_n$, we obtain since $\psi_n\in[-M,M]$

$$t_n\psi_n.x'(0)+o(t_n)+t_nh_n(t_n\psi_n-)\leqslant 0\leqslant \tag{16}$$
$$\leqslant t_n\psi_n.x'(0)+o(t_n)+t_nh_n(t_n\psi_n)$$

As $n\to\infty$, $t_n\psi_n\to 0$, so by uniform convergence of h_n to h and continuity of h at 0 we obtain on dividing (16) throughout by t_n

$$(\text{lim sup } \psi_n).x'(0)+h(0)\leqslant 0\leqslant(\text{lim inf } \psi_n).x'(0)+h(0)$$

or lim $\psi_n=-h(0)/x'(0)$.

It remains to establish that lim sup$|\psi_n|<\infty$. Now because $x'(0)>0$ and x is nondecreasing $\exists a>0$ and $c>0$ such that (see Figure 2)

$$x(u)\geqslant cu \quad 0\leqslant u\leqslant a$$

$$x(u)\geqslant ca \quad a\leqslant u$$

Let $A<\infty$ be an upper bound to $|h_n|$ on $\overline{\mathbb{R}}$ for all n. Then

$$x_n(u) = x(u)+t_nh_n(u)\geqslant\begin{cases} cu-t_nA & 0\leqslant u\leqslant a \\ ca-t_nA & a\leqslant u\end{cases}$$

Thus if n is sufficiently large that $ca-t_nA>0$, we have $x_n(u)>0$ for $u>t_nA/c$. Similarly $x_n(u)<0$ for $u<-t_nA/c$ for large enough n. Since $x_n(\theta_n-)\leqslant 0\leqslant x_n(\theta_n)$ we must have $|\theta_n|\leqslant t_nA/c$ for large enough n and hence lim sup$|\psi_n|\leqslant A/c<\infty$. □

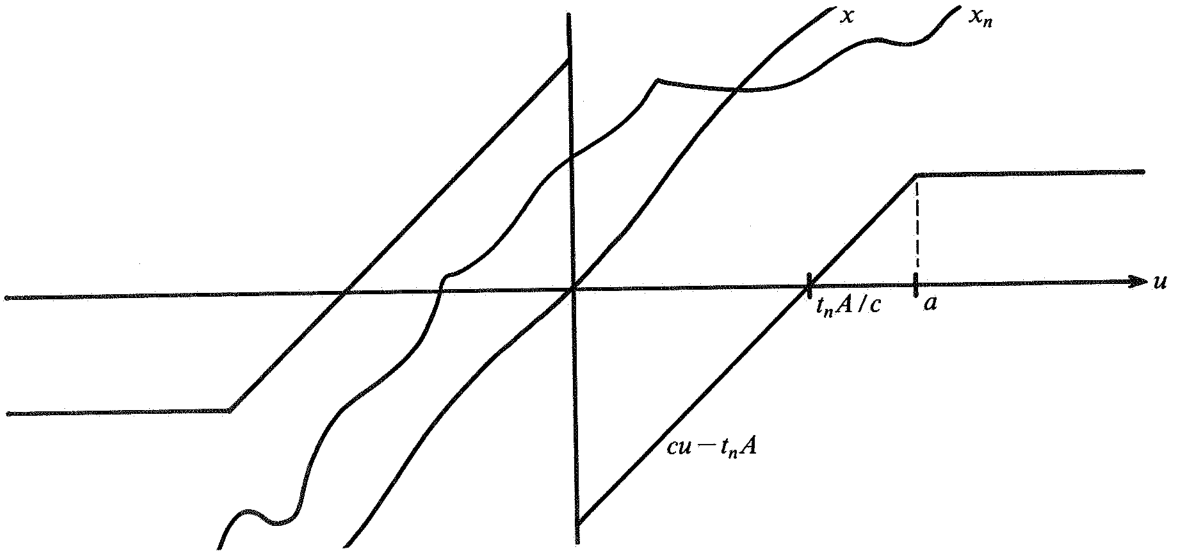

FIGURE 2. Proof of $\limsup|\psi_n|<\infty$.

Taking Lemma 1 of Section 2.2. into account, we obtain the following corollaries, where $E\subset D(\mathbb{R})$ is the set of distribution functions on $\mathbb{R}$.

COROLLARY 2. *Let $x\in E$ be such that x is differentiable at the point $\theta\in(0,1)$, and $x(\theta)=p$, $x'(\theta)>0$. Suppose $\phi:E\to[0,1]$ satisfies*

$$y(\phi(y)-)\leq p\leq y(\phi(y))$$

for all y in some neighbourhood of x. Then ϕ can be extended to $D(\overline{\mathbb{R}})$ so as to be Hadamard differentiable at x tangentially to $\{h\in D(\overline{\mathbb{R}}):h$ is continuous at $\theta\}$, with derivative

$$\mathrm{d}\phi(x).h = -h(\theta)/x'(\theta).$$

Note that the derivative is indeed a continuous linear map from $(D(\overline{\mathbb{R}}),\|.\|)$ to $\mathbb{R}$. Combining this with Corollary 1 gives

COROLLARY 3. *If T_n is a p'th quantile of an empirical distribution function F_n based on a random sample of size n from a distribution F with $F(\theta)=p$, F differentiable at θ with $f(\theta) = F'(\theta)>0$, then*

$$n^{\frac{1}{2}}(T_n-\theta)\xrightarrow{\mathfrak{D}}-\frac{B^0(F(\theta))}{F'(\theta)}\overset{\mathfrak{D}}{=}N\left[0,\frac{p(1-p)}{f(\theta)^2}\right].$$

Though we have restricted attention to a quantile of an empirical distribution function based on a random sample, the method of proof applies to obtaining the limiting distribution of an inverse of any one-dimensional empirical process: we just need (e.g.) continuous sample paths of the limiting process. Also the method can be extended to give, via differentiability of the mapping $F\to F^{-1}$, weak convergence of the whole quantile process (for a similar approach see VERVAAT (1972)). We summarize the differentiability result, without proof, as a variant of Corollary 2:

PROPOSITION 1. *Let x in E be continuously differentiable with positive derivative on an interval*

$[a-\epsilon,b+\epsilon]$, $\epsilon>0$, $a<b$. For $y\in E$, let $\phi(y)\in D[p_0,p_1]$ denote the right-continuous inverse of y, restricted to the interval with endpoints $p_0=x(a)$, $p_1=x(b)$. Then ϕ can be extended so as to be compactly differentiable at x, tangentially to the space $\{h\in D(\overline{\mathbb{R}}):h$ is continuous on $[a-\epsilon,b+\epsilon]\}$ with derivative $d\phi(x).h = -\left[\frac{h}{x'}\right]\circ x^{-1}$ mapping $D[a,b]$ to $D[p_0,p_1]$.

In our second example we are again confronted with the problem of extending a functional of empirical distribution functions to $D(\overline{\mathbb{R}})$. We omit the details of the application to the Wilcoxon statistic, but just recall that this can be constructed from the mapping $(F,G)\rightarrow\int_{-\infty}^{\infty}FdG$ for two distribution functions F and G. We shall investigate the differentiability of this rather simple mapping. Surprisingly this is not a trivial matter. More generally, one is usually only interested in the mapping $\phi(x,y)=\int_{-\infty}^{\infty}xdy$ for elements x and y of $D(\overline{\mathbb{R}})$ which are both of bounded variation. In fact, we can often without loss of generality suppose that the total variation of the sample counterpart of y (and indeed of x too) is bounded by a finite constant (with probability converging to 1 as $n\rightarrow\infty$). This motivates the choice of a subset E of $D(\overline{\mathbb{R}})^2$ to which we will restrict ϕ:

LEMMA 3. *Let $E=\{(x,y)\in D(\overline{\mathbb{R}})^2:\int_{-\infty}^{\infty}|dy|\leqslant C\}$ and let $\phi:E\rightarrow\mathbb{R}$ be defined by $\phi(x,y)=\int_{-\infty}^{\infty}xdy$. Let (x,y) now be a fixed point of E such that $\int_{-\infty}^{\infty}|dx|$ is finite too. Then ϕ can be extended to $D(\overline{\mathbb{R}})^2$ so as to be Hadamard differentiable at (x,y), with derivative*

$$d\phi(x,y)\cdot(h,k) = \int_{-\infty}^{\infty}xdk + \int_{-\infty}^{\infty}hdy,$$

where the integral with respect to k is defined by the integration by parts formula if k is not of finite variation.

Here we interpret integration from $-\infty$ to ∞ as being over the interval $(-\infty,\infty]$. The integration by parts formula, for $a,b\in\overline{\mathbb{R}}$, is then

$$\int_{(a,b]}x\,dk = x(b)k(b)-x(a)k(a) - \int_{(a,b]}k_-dx$$

where k_- is the left-continuous modification of k. Other conventions as to the range of integration need a corresponding modification of this formula. Note also that the lemma does not assert that a single extension exists, differentiable at *all* points in E!

PROOF. Suppose the sequences $t_n\in\mathbb{R}^+, h_n\in D(\overline{\mathbb{R}})$, and $k_n\in D(\overline{\mathbb{R}})$ satisfy $t_n\rightarrow 0$, $h_n\overset{\|\cdot\|}{\rightarrow}h$ and $k_n\overset{\|\cdot\|}{\rightarrow}k$. Define $x_n=x+t_nh_n$, $y_n=y+t_nk_n$ for given x and $y\in D(\overline{\mathbb{R}})$ and suppose $(x_n,y_n)\in E$ for each n.

First we note that the hypothesized derivative *is* a continuous linear map; i.e.

$$(\int xdk_n, \int h_ndy)\rightarrow(\int xdk, \int hdy) \text{ as } n\rightarrow\infty,$$

where we have dropped the range of integration. Now we have

$$t_n^{-1}(\phi(x_n,y_n) - \phi(x,y)) - \int xdk_n - \int h_ndy = t_n\int h_ndk_n = \int h_nd(y_n-y).$$

It therefore suffices to show that $\int h_nd(y_n-y)\rightarrow 0$ as $n\rightarrow\infty$; and since

$$|\int(h_n-h)d(y_n-y)|\leqslant(\|h_n - h\|)(\int|dy_n| + \int|dy|)$$

it suffices to show $\int hd(y_n-y)\rightarrow 0$.

For any $\epsilon>0$ there exists $h'\in D(\overline{\mathbb{R}})$ such that h' is a step function with a finite number (say N) of jumps and $\|h'-h\|\leqslant\epsilon$. We have

$$|\int(h-h')d(y_n-y)|\leqslant\epsilon(\int|dy_n|+\int|dy|)\leqslant 2\epsilon C$$

while, by partial integration,

$$\begin{aligned}|\textstyle\int h'\mathrm{d}(y_n-y)| &\leqslant 2\|h'\|\cdot\|y_n-y\| \\ &\quad + \|y_n-y\|\textstyle\int|\mathrm{d}h'| \\ &\leqslant 2\|h'\|\cdot\|y_n-y\| + 2N\|h'\|\cdot\|y_n-y\| \\ &\to 0 \text{ as } n\to\infty.\end{aligned}$$

Therefore

$$\limsup|\textstyle\int h\mathrm{d}(y_n-y)|\leqslant 2\epsilon C$$

and since ϵ was arbitrary, the result is proved. □

This proof actually shows that the mapping $(x,y)\to\int_{-\infty}^{(\cdot)}x\mathrm{d}y$, from E to $D(\overline{\mathbb{R}})$, can be extended so as to be differentiable at a point satisfying the hypothesis of the lemma; the derivative is the same with $\int_{-\infty}^{\infty}$ replaced by $\int_{-\infty}^{(\cdot)}$. This can be applied in the many examples from survival analysis (e.g. estimation of a cumulative hazard rate, k-sample tests) which involve this functional.

Applied to the Wilcoxon test, we obtain a perfectly general result on the asymptotic normality of $\int F_n\mathrm{d}G_m$ (i.e. without any continuity restrictions on the underlying distribution functions F and G).

However there is also a negative aspect to Lemma 3. The functional ϕ is only differentiable at a point in $(D(\overline{\mathbb{R}}))^2$ satisfying finite variation properties, and is clearly not differentiable in a whole *neighbourhood* of such a point, and certainly not continuously differentiable. One can easily exhibit sequences $x_n\to x$, $k_n\to k$ such that $\int x_n\mathrm{d}k_n\not\to\int x\mathrm{d}k$. So the implicit function theorem cannot be applied to proving existence and differentiability of solutions to equations involving this functional; at least not with the present choice of topology on $D(\overline{\mathbb{R}})$.

The proof of Lemma 3 is actually just the usual proof of the Helly-Bray Lemma; see in particular the version called Helly's theorem in SMIRNOV (1972). One can see exactly the same proof being carried out in a statistical context in BRESLOW & CROWLEY (1974) and in many other papers. Perhaps one can say that the contribution of Hadamard differentiability in such a context is simply to show that what is being done is just a verification of differentiability, for instance the whole proof of Breslow & Crowley truly is “just” an application of the δ-method. One can complete a von Mises treatment of the product-limit estimator by proving Hadamard differentiability of the functional $x\to\prod_0^{(\cdot)}(1+\mathrm{d}x)$; see GILL & JOHANSEN (1987). Also, these few examples may appear quite complicated, but once one has established differentiability of a few key functionals, the chain rule yields differentiability of a huge class of composite functionals and the elegance of the approach becomes apparent. For an amusing application see GRÜ'BEL (1988).

We feel that the extension lemma (Lemma 1), the notion of differentiability tangentially to a subspace, and the use of POLLARD'S (1984) weak convergence theory, considerably streamline REEDS' (1976) original approach. He for instance used the equivalent definition (5) and outer probability arguments rather than (4) and the Skorohod-Dudley-Wichura representation theorem. Other measurability and differentiability problems are solved by Reeds, Fernholz and more lately ESTY et al. (1985) and TAYLOR (1985) by constructing continuous modifications of empirical distribution functions and then working in $C[0,1]$. This means that their theorems apply in the first place to approximations of the original statistics of interest, and are only applicable when the underlying distribution function F is continuous. Justification of all these ad hoc approximations distracts from the simplicity of the basic δ-method.

2.4. The δ-method and the bootstrap

Bootstrapping involves taking random samples from the empirical distribution F_n. Thus if we want to show that "the bootstrap works" for estimating the distribution of a statistic $\phi(F_n)$ by calculating the distribution of $\phi(F_n^*)$ given F_n, where F_n^* is the empirical distribution based on a sample of size n from F_n, it seems as though we need smoothness of ϕ not just at F but also in its neighbourhood. For instance, PARR (1985a,b) uses the *continuous* (Frechet) differentiability of some functionals, following REEDS (1978) and BICKEL & FREEDMAN (1981), to treat jackknife and bootstrap. (CLARKE (1983, 1984) and REEDS (1976) show that M-estimators have these properties). See also YANG (1985). However, just Hadamard differentiability at a point, or a very weak form of continuous Hadamard differentiability, are sufficient to get some weak bootstrap results. Here we briefly give some of these; related work has been done by LOHSE (1988) and LIU, SINGH & LO (1986). Essentially we show that 'the bootstrap works if the δ-method works'; both give asymptotically the same answer. Of course it is often the case that the bootstrap works better, but our theory says nothing on this matter.

We only consider the case when the X_i's are real so that (for our first result) the 'probability transform' works: $X_i \overset{\mathcal{D}}{=} F(U_i)$ where U_i is uniform [0,1]; the idea of using this for the bootstrap comes from SHORACK & WELLNER (1985, Ch. 2.3.). Let F_n be the empirical d.f. based on $X_1,\ldots,X_n$, the first n of an infinite sequence of real i.i.d. random variables with distribution function F. Let $B_1 = (D(\overline{\mathbb{R}}), \|\cdot\|)$ endowed with the open-ball σ-algebra. Let F_n^* be a bootstrap sample of size n from F_n.

THEOREM 4. (Strong consistency of the bootstrap under weak continuous compact differentiability). *Suppose* $\phi : B_1 \to B_2$ *is measurable and continuously compact differentiable at* $x = F$, *in the sense that* $x_n \xrightarrow{\|\cdot\|} x$, $h_n \xrightarrow{\|\cdot\|} h$, $t_n \to 0 \in \mathbb{R}$ *implies*

$$\frac{\phi(x_n + t_n h_n) - \phi(x_n)}{t_n} \to \mathrm{d}\phi(x)\cdot h$$

where $\mathrm{d}\phi(x) : B_1 \to B_2$ *is continuous, and linear. Then*

$$n^{\frac{1}{2}}(\phi(F_n^*) - \phi(F_n)) \xrightarrow{\mathcal{D}} \mathrm{d}\phi(F)\cdot B^0 \circ F \ \textit{a.s.}$$

Note that this is also the limiting distribution of $n^{\frac{1}{2}}(\phi(F_n - \phi(F))$.

PROOF. We have a.s. $F_n \xrightarrow{\|\cdot\|} F$. Moreover, given F_n

$$n^{\frac{1}{2}}(F_n^* - F_n) \overset{\mathcal{D}}{=} n^{\frac{1}{2}}(U_n^* - \iota)\circ F_n$$

where U_n^* is the empirical d.f. based on n i.i.d. uniform [0,1] r.v.'s and ι is the identity mapping. Now

$$n^{\frac{1}{2}}(U_n^* - \iota) \xrightarrow{\mathcal{D}} B^0 \text{ in } D([0,1]), \|\cdot\|.$$

By the Skorohod-Dudley-Wichura representation theorem there exist $U_n^{*\prime} \overset{\mathcal{D}}{=} U_n^*$, $B^{0\prime} \overset{\mathcal{D}}{=} B^0$, with

$$n^{\frac{1}{2}}(U_n^{*\prime} - \iota) \xrightarrow{\|\cdot\|} B^{0\prime} \text{a.s.}$$

Define $F_n^{*\prime} = U_n^* \circ F_n$. Given $F_1, F_2, \ldots$ s.t. $\|F_n - F\| \to 0$, we have

$$n^{\frac{1}{2}}(F_n^* - F_n) \overset{\mathcal{D}}{=} n^{\frac{1}{2}}(F_n^{*\prime} - F_n).$$

Moreover by continuity of the paths of $B^{0\prime}$,

$$n^{\frac{1}{2}}(F_n^{*\prime} - F_n) = n^{\frac{1}{2}}(U_n^{*\prime} - \iota)\circ F_n \xrightarrow{\|\cdot\|} B^{0\prime}\circ F.$$

By the definition of continuous compact differentiability of ϕ, and the assumption $\|F_n - F\|\to 0$,

$$n^{\frac{1}{2}}(\phi(F_n^{*\prime}) - \phi(F_n))\to \mathrm{d}\phi(F)\cdot B^{0\prime}\circ F.$$

Therefore

$$n^{\frac{1}{2}}(\phi(F_n^{*}) - \phi(F_n))\xrightarrow{\mathfrak{D}} \mathrm{d}\phi(F)\cdot B^{0}\circ F$$

on the event $\{\|F_n - F\|\to 0\}$; i.e. almost surely. □

The above theorem used the ‘probability transform’. For the next result we also use a technique specific to $\mathbb{R}^1$, but less strong, so that there is hope for an extension. Suppose also ϕ is now only ordinarily compact differentiable at $x = F$; i.e.

$$\frac{\phi(x+t_n h_n)-\phi(x)}{t_n}\to \mathrm{d}\phi(x)\cdot h$$

for $h_n \xrightarrow{\|\cdot\|} h$, $t_n \to 0 \in \mathbb{R}$. We would like to prove that

$$\mathfrak{L}(n^{\frac{1}{2}}(\phi(F_n^{*}) - \phi(F_n)))\xrightarrow{P}\mathfrak{L}(\mathrm{d}\phi(F)\cdot B^{0}\circ F)$$

when $\mathfrak{L}(\cdot)$ means: “the distribution of”. To make sense of this we need to metricize Pollard’s more general weak convergence concept; this can be done using the *bounded-Lipschitz* metric; Pollard (1984, p. 74). However we prefer not to introduce that here. Since for some applications we would like ϕ to take values in a non-separable space, e.g. $D(\overline{\mathbb{R}}), \|\cdot\|$, it would also be undesirable to restrict attention to separable B_2, for which the well-known Lévy-Prohorov metric

$$d(P_1,P_2) = \inf\{\epsilon : P_1(A)\leqslant P_2(A^{\epsilon})+\epsilon \ \forall A\in\mathfrak{B}\}$$

metricizes weak convergence (here A^{ϵ} is the union of all open ϵ-balls around points of A). A sensible compromise is to look at functionals $\psi(n^{\frac{1}{2}}(\phi(F_n^{*}) - \phi(F_n)))$, where ψ maps B_2 into a separable space, e.g. $\mathbb{R}^1$. Note that on $\mathbb{R}^1$,

$$\begin{aligned} d(P_1,P_2) = \inf\{\epsilon : {}& P_1(-\infty,t]\leqslant P_2(-\infty,t+\epsilon)+\epsilon,\\ & P_2(-\infty,t]\leqslant P_1(-\infty,t+\epsilon)+\epsilon \ \ \forall t\}\end{aligned}$$

and if $P(-\infty,t]$ is continuous in t, then

$$d(P^n,P)\to 0 \Leftrightarrow \sup_t\{|P^n(-\infty,t]-P(-\infty,t]|\}\to 0.$$

Theorem 5. (Weak consistency of bootstrap). *Let F_n, F_n^{*}, ϕ be as before, except now ϕ is only required to be ordinarily compact differentiable at $x = F$. Let $\psi : B_2\to\mathbb{R}$ be measurable and continuous at all points of a subset of B_2 in which $\mathrm{d}\phi(F)\cdot B^0\circ F$ lies with probability one. Then*

$$d(\mathfrak{L}^{*}(\psi(n^{\frac{1}{2}}(\phi(F_n^{*})-\phi(F_n)))),\mathfrak{L}(\psi(\mathrm{d}\phi(F)\cdot B^0\circ F)))\xrightarrow{P} 0$$

as $n\to\infty$. In particular, if $\psi(\mathrm{d}\phi(F)\cdot B^0\circ F)$ has a continuous distribution,

$$\sup_t|P^{*}[\psi(n^{\frac{1}{2}}(\phi(F_n^{*}) - \phi(F_n)))\leqslant t] - P[\psi(\mathrm{d}\phi(F)\cdot B^0\circ F)\leqslant t]|\xrightarrow{P} 0$$

as $n\rightarrow\infty$. ($\mathcal{L}^*, P^*$ denote bootstrap distributions, for fixed F_n).

Proof. First use the Skorohod-Dudley-Wichura representation theorem to construct a sequence $F_n' \overset{\mathcal{D}}{=} F_n$ with $n^{\frac{1}{2}}(F_n'-F)\xrightarrow{\|\cdot\|}Z'$ a.s., where $Z'\overset{\mathcal{D}}{=}Z=B^0\circ F$. Let $F_n'^*$ be a bootstrap sample empirical d.f. of size n from F_n'. As before, we have $n^{\frac{1}{2}}(F_n'^*-F_n')\xrightarrow{\mathcal{D}}Z^*\overset{\mathcal{D}}{=}Z'$. By Skorohod-Dudley-Wichura again, (for a given sequence F_n', with $n^{\frac{1}{2}}(F_n'-F)\xrightarrow{\|\cdot\|}Z'$) we can construct $F_n'^{*\prime}\overset{\mathcal{D}}{=}F_n'^*$ with

$$n^{\frac{1}{2}}(F_n'^{*\prime} - F_n')\xrightarrow{\|\cdot\|}Z^{*\prime}\overset{\mathcal{D}}{=}Z^*\text{a.s.}$$

Thus

$$n^{\frac{1}{2}}(F_n'^{*\prime}-F)\xrightarrow{\|\cdot\|}Z^{*\prime}+Z'$$

and

$$n^{\frac{1}{2}}(F_n'-F)\xrightarrow{\|\cdot\|}Z'.$$

So

$$\begin{aligned} n^{\frac{1}{2}}(\phi(F_n'^{*\prime})-\phi(F_n')) &= n^{\frac{1}{2}}(\phi(F_n'^{*\prime}) - \phi(F)) - n^{\frac{1}{2}}(\phi(F_n') - \phi(F)) \\ &\rightarrow \mathrm{d}\phi(F)\cdot(Z^{*\prime}+Z') - \mathrm{d}\phi(F)\cdot Z' \\ &= \mathrm{d}\phi(F)\cdot Z^{*\prime} \end{aligned}$$

and hence, by continuity of ψ, $\mathrm{d}\phi(F)\cdot Z-$ a.s.

$$\psi(n^{\frac{1}{2}}(\phi(F_n'^{*\prime}) - \phi(F_n')))\rightarrow\psi(\mathrm{d}\phi(F)\cdot Z^{*\prime})\text{ a.s.}$$

Since

$$\psi(n^{\frac{1}{2}}(\phi(F_n'^{*\prime})-\phi(F_n')))\overset{\mathcal{D}^*}{=}\psi(n^{\frac{1}{2}}(\phi(F_n'^{*})-\phi(F_n')))$$

(equality of bootstrap distributions -F_n' is fixed at present) we have for our sequence F_n', a.s.,

$$\psi(n^{\frac{1}{2}}(\phi(F_n'^{*}) - \phi(F_n')))\xrightarrow{\mathcal{D}^*}\psi(\mathrm{d}\phi(F)\cdot Z)$$

Thus

$$d(\mathcal{L}^*[\psi(n^{\frac{1}{2}}(\phi(F_n'^{*}) - \phi(F_n')))],\mathcal{L}[\psi(\mathrm{d}\phi(F)\cdot Z)])\rightarrow 0$$

a.s. . One can check that the left-hand side here is a measurable function of $F_n'\overset{\mathcal{D}}{=}F_n$. Thus we may conclude

$$d(\mathcal{L}^*[\psi(n^{\frac{1}{2}}(\phi(F_n^{*}) - \phi(F_n)))],\mathcal{L}[\psi(\mathrm{d}\phi(F)\cdot Z)])\xrightarrow{P}0.$$

The further result for the case when $\psi(\mathrm{d}\phi(F)\cdot Z)$ is continuously distributed is obtained similarly. □

Note 1. This proof used the fact that if F_n is a sequence of fixed non-random distribution functions on $\mathbb{R}^1$, with $n^{\frac{1}{2}}(F_n-F)\xrightarrow{\|\cdot\|}Z$, and F_n^* is the empirical based on a sample of size n from F_n, then $n^{\frac{1}{2}}(F_n^*-F_n)$ converges in distribution just as if F_n had not depended on n. So the proof can be extended to $\mathbb{R}^p$ and other norms, provided this implication still holds; i.e. that the bootstrap applied to the empirical process itself works. See GAENSSLER (1987) for results in this direction.

NOTE 2. The same proof shows that, under the same hypotheses on F_n and ϕ,

$$\lambda(\mathcal{L}^*(n^{\frac{1}{2}}(\phi(F_n^*)-\phi(F_n))),\mathcal{L}(\mathrm{d}\phi(F)\cdot B^0\circ F))\xrightarrow{P}0$$

where λ is Pollard's bounded Lipschitz metric for weak convergence in his more general sense,

$$\lambda(P,Q) = \sup\{|E_P f-E_Q f|:\text{bounded, measurable } f \text{ with } |f(x)-f(y)|\leqslant\|x-y\|\ \forall x,y\}.$$

NOTE 3. Typical examples for these theorems would be: the Kaplan-Meier median (only differentiable at a point, targentially to a subspace; and not continuously differentiable); or taking ψ to be the supremum norm of a function in $D(\overline{\mathbb{R}})$, one can apply them to constructing bootstrap confidence bounds for such empirical processes as the total time on test plot or the residual quantile lifetime function (based on the Kaplan-Meier estimator), again only compactly differentiable at a point. (Differentiability *tangentially* to a certain subspace is easily incorporated in the theorem). In all these examples we work in the usual random censorship model, for which we can easily embed the bivariate observations (X_i,δ_i), $\delta_i=0$ or1, into the real line.

3. NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATION

Much literature is devoted to discussions of how a nonparametric maximum likelihood estimator (NPMLE) should be defined; see especially KIEFER & WOLFOWITZ (1956), SCHOLZ (1980), JOHANSEN (1978, 1983), and JACOBSEN (1984). (We do *not* discuss here the alternative ways of adapting the maximum likelihood principle employed in the method of sieves, GRENANDER (1981); or the method of penalized likelihood, see GEMAN & HWANG (1982) for a comparison of these two principles). From the point of view of large sample theory these discussions have been, at least till now, irrelevant: there is also no theory of large sample properties of NPMLE's which is relevant to any interesting practical examples.

Two points are central in these discussions. Firstly, since typically no dominating measure exists in such problems (think of the problem of estimating an arbitrary unknown distribution function F) one cannot define the NPMLE by just "maximizing a density". Kiefer & Wolfowitz's approach is to consider pairwise comparisons only. So we say that $\hat{\alpha}$ is an MLE based on data X from the model $\{P_\alpha:\alpha\in\mathcal{Q}\}$, where $\mathcal{Q}$ may be infinite-dimensional and P_α is the distribution of X on the sample space $\mathcal{X}$ under α, if

$$\frac{\mathrm{d}P_{\hat{\alpha}}}{\mathrm{d}\mu}(X)\geqslant\frac{\mathrm{d}P_\alpha}{\mathrm{d}\mu}(X)$$

for all $\alpha\in\mathcal{Q}$ and $\mu\gg P_{\hat{\alpha}}$, P_α; so we take a different μ — e.g. $P_\alpha+P_{\alpha'}$ — when comparing each α, $\alpha'\in\mathcal{Q}$ (Scholz addresses the problem that $\mathrm{d}P_\alpha/\mathrm{d}\mu$ is only defined μ-a.e., so this definition depends on an arbitrary choice of versions of Radon-Nikodym derivatives).

Secondly, even with this sensible definition, an MLE often just does not exist. Consider for example the model: $X_1,\ldots,X_n$ is a random sample from a *continuous* distribution F. The empirical distribution function F_n should be the NPMLE, but unfortunately it is discrete and hence not in the parameter space. In such a simple example an obvious *discrete extension* of the original *continuous* model exists. However in more complicated models for an essentially continuous phenomenon — e.g. COX'S (1972) regression model — several different discrete extensions of the model can be constructed, each a natural extension from some point of view, but each leading to a *different* NPMLE. Typically, at an underlying "continuous" point in the model, the different estimators are asymptotically equivalent. See JOHANSEN (1983) and JACOBSEN (1984) for some examples of this.

Our approach suggests that this search for "the correct discrete extension" of a given continuous model has been addressing the wrong criteria. If one is interested in NPMLE's because of their hopefully good asymptotic properties *at a point in the original model*, one should try to extend *score functions* (or likelihood equations) from continuous to discrete points in the parameter space in as smooth a way as possible, in particular so as to obtain differentiability at an underlying continuous point in the model. One must be able to approximate a continuous point arbitrarily well with discrete ones, not vica-versa. The extended score function at a point α in the extended parameter space need not even correspond to an actual model — i.e. a distribution P_α — for the observations X.

We shall return to this second point later. For the time being, we will follow the Kiefer-Wolfowitz definition of an MLE and suppose that our parameter space is large enough that it exists. By means of examples, we show that the NPMLE is often determined as the solution of the likelihood equations for a collection of smooth parametric submodels. These equations are in fact precisely the "self-consistency" equations introduced by Efron (1967) and more recently studied, using von Mises methods based on Frechet differentiability, by Crowley & Tsai (1985).

Suppose we have data X coming from some model $\{P_\alpha : \alpha \in \mathcal{Q}\}$ where the parameter space $\mathcal{Q}$ is some large (i.e. infinite dimensional) collection of e.g. distribution functions, cumulative hazard functions, or pairs, each consisting of such an object together with a Euclidean parameter. Our claim is that in many such examples, one can construct mappings $\phi(\alpha,h,\theta) \in \mathcal{Q} : \alpha \in \mathcal{Q}, h \in H, \theta \in \mathbb{R}$ such that $\phi(\alpha,h,0)=\alpha$ for all h. Thus for each $\alpha \in \mathcal{Q}$ and $h \in H$, the model $\{P_{\phi(\alpha,h,\theta)} : \theta \in \mathbb{R}\}$ is a one-dimensional parametric submodel of the original model, which passes (at $\theta=0$) through the point P_α. Here H can sometimes be interpreted as a set of directions, or as indexing the possible directions with which such a parametric sub-model passes through the point P_α. Later (in Part II) we also consider two-dimensional parametric submodels generated by mappings $\phi(\alpha,h,k;\theta,\psi)$ within which our one-dimensional submodels are nested: $\phi(\alpha,h,\theta)=\phi(\alpha;h,k;\theta,0)=\phi(\alpha;k,h;0,\theta)$ for all α,h,k,θ.

Now if $\{P_{\phi(\alpha,h,\theta)} : \theta \in \mathbb{R}\}$ is a dominated family of probability measures for each α and h, if the corresponding density is a differentiable function of θ for all $x \in \mathcal{X}$, and if an NPMLE $\hat{\alpha}=\hat{\alpha}(X)$ exists, then we must have:

$$U_h(\hat{\alpha};X) = 0 \quad \text{for all } h \in H \tag{19}$$

where

$$U_h(\alpha;X) = \frac{\partial}{\partial\theta} \log \text{lik}\,(\theta,X;\alpha,h)|_{\theta=0} \tag{20}$$

and

$$\text{lik}\,(\theta;x;\alpha,h) = \frac{\mathrm{d}P_{\phi(\alpha,h,\theta)}}{\mathrm{d}\mu}(x) \tag{21}$$

for a suitably chosen dominating measure $\mu=\mu(\alpha,h)$. In many examples $\hat{\alpha}(X)$ is actually uniquely determined by the equations (19).

In other examples, modelling a continuous phenomenon, an NPMLE according to the Kiefer-Wolfowitz criterium may not exist and correspondingly (19) may not have a solution. However it often then happens that the function $U_h(\alpha;X)$ can be extended in a natural way from $\alpha \in \mathcal{Q}$ to $\alpha \in \overline{\mathcal{Q}}$ for some larger set $\overline{\mathcal{Q}}$, on which (19) *does* have a solution.

Let us illustrate these ideas by a series of examples.

Example 1. *The empirical distribution function.*

Suppose $X_1, \ldots, X_n$ are a random sample from some distribution function F on $\mathbb{R}^d$, which is completely unknown. So we identify the parameter α with F and the parameter space $\mathcal{Q}$ with $\mathcal{F}$, the set of all d.f.'s on $\mathbb{R}^d$. Let H be the space of all bounded measurable functions on $\mathbb{R}^d$. For any d.f. F, any $h \in H$, and for all $\theta \in \mathbb{R}^1$ sufficiently close to 0, define a distribution function $\phi(F,h,\theta)$ absolutely continuous with respect to F by

$$\frac{\mathrm{d}\phi}{\mathrm{d}F}(F,h,\theta) = \frac{1+\theta h}{\int (1+\theta h)\mathrm{d}F}\,.$$

Then the distribution of $X=(X_1,\ldots,X_n)$ under $\phi(F,h,\theta)$ is dominated by its distribution under F itself, with Radon-Nikodym derivative

$$\mathrm{lik}\ (\theta;X;F,h) = \prod_{i=1}^{n}\frac{1+\theta h(X_i)}{\int(1+\theta h)\mathrm{d}F}$$

So

$$\log \mathrm{lik}\ (\theta;X;F,h)=\sum\log(1+\theta h(X_i)) - n\ \log\int(1+\theta h)\mathrm{d}F$$

and

$$\begin{aligned} U_h(F;X) &= \frac{\partial}{\partial\theta}\log \mathrm{lik}\ (\theta;X;F,h)|_{\theta=0} \\ &= \sum h(X_i)-n\int h\mathrm{d}F \\ &= n\int h\mathrm{d}(F_n-F)=n\int(h-\int h\mathrm{d}F)\mathrm{d}F_n \end{aligned}$$

where F_n is the empirical distribution function based on $X_1,\ldots,X_n$. So the likelihood equations (19) reduce to

$$n\int h\mathrm{d}(F_n-\hat{F}) = 0 \quad \forall h\in H \tag{22}$$

which has the unique solution $\hat{F}=F_n$. In fact H could have been reduced to the collection of quadrant indicator functions $1_{(-\infty,x]}, x\in\mathbb{R}^d$, in which case (19) becomes

$$n(F_n(x) - \hat{F}(x))=0 \quad \forall x\in\mathbb{R}^d \tag{23}$$

Typically we will find that the likelihood equations can be reduced to a collection "of the same dimension" as the parameter space $\mathcal{A}$. In the i.i.d. case it is always so that the likelihood equations depend on the data through its empirical d.f., moreover the dependence is linear. Thus considering U_h for each $h\in H$ as the component of a vector (or evaluation of a function) U, we rewrite (19) as

$$nU(\hat{\alpha}_n,F_n) = 0 \tag{24}$$

where U maps $\mathcal{A}\times\{\text{empirical d.f.'s }\}$ to a new space of similar structure to $\mathcal{A}$, and where U is linear in F_n. Under the usual interchange (if valid) of expectation and integration, the expected side of the left hand sides of (19) and (24) are zero and we have *Fisher consistency* of the NPMLE $\hat{\alpha}_n$: letting F_α denote the d.f. of one observation under P_α, we have $U(\alpha,F_\alpha)=0$.

EXAMPLE 2 *Grouped and censored data from an unknown distribution*

Continuing EXAMPLE 1, suppose we do not actually observe the random sample $X_1,\ldots,X_n$ itself, but only some many-to-one function of this sample. For instance, we might only observe for each i the pair $(X_i 1_{B_i}(X_i), 1_{B_i}(X_i))$ where $B_i\subseteq\mathbb{R}^d$ are known (non-random) sets, e.g. intervals. Thus for each i the value of X_i is observed if it falls in B_i, otherwise one only observes the occurrence of the event "$X_i\notin B_i$". In the case $d=1$, if $B_i=(-\infty,a_i]$ for each i and some constants $a_i\in\mathbb{R}$, this is the familiar model of (fixed) right censoring. More general specifications lead to general models for grouped or censored data. (For instance, with bivariated censored data, the B_i's have to be taken as random sets determined by the underlying survival times and censoring times jointly). TURNBULL (1976) discusses an estimator of the underlying d.f. F of the X_i's based on grouped or censored data which in the model with the B_i's is defined as the limit, if it exists, of the iterations:

$$F^{(k+1)}(x) = \frac{1}{n}\sum_i\begin{cases} 1_{(-\infty,x]}(X_i) & \text{if } X_i \text{ is observed} \\ E_{F^{(k)}}\{1_{(-\infty,x]}(X^*)|X^*\notin B_i\} & \text{if } X_i \text{ is not observed}\end{cases} \tag{25}$$

Here X^* is drawn from the distribution $F_{(k)}$, the current estimate of F at the k'th iteration. This simple algorithm has great intuitive appeal and can be considered as the application of the EM-algorithm (Dempster, Laird & Rubin, 1977) to this problem. However almost nothing is known about large-sample properties of the resulting estimator except in some very special situations (e.g. the $d=1$, right censoring case, when we obtain the well-known product-limit estimator as limit provided a sensible initial choice $F^{(0)}$ is made; and the double censoring problem, Chang & Yang, 1988)

We can relate the algorithm directly to the score equation (23) of Example 1, and to Efron's (1967) self-consistency principle, as follows. Let $X=(X_1,\ldots,X_n)$ be the not completely observable underlying sample from F, and let $Y=g(X)$ be the observable data where g is some many-to-one map. Consider a parametric submodel in which Y has density $f_Y(y;\theta)$ too. Usually we will then have

$$\frac{\partial}{\partial\theta}\log f_Y(y;\theta) = E_\theta(\frac{\partial}{\partial\theta}\log f_X(X;\theta)|Y=y)\,. \tag{26}$$

To confirm this, note that for $y=g(x)$ we have

$$f_X(x;\theta) = f_Y(y;\theta)f_{X|Y=y}(x;\theta)\,.$$

So taking logarithms, differentiating with respect to θ, substituting X for x, and finally taking expectations with respect to the conditional distribution of X given $Y=y$, we obtain (26) since if the usual interchange of iteration and differentiation is valid,

$$E_\theta\left[\frac{\partial}{\partial\theta}\log f_{X|Y=y}(X;\theta)\middle|Y=y\right] = 0\,.$$

Thus for the parametric submodel of Example 1,

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\log \text{lik}(\theta;Y;F,h)|_{\theta=0}\\
&\quad= E_\theta\left[\frac{\partial}{\partial\theta}\log \text{lik}(\theta;X;F,h)\middle|Y\right]|_{\theta=0}\\
&\quad= E_F\left[n\int h\,\mathrm{d}(F_n-F)\middle|Y\right] \quad \text{since } \phi(F,h,0)=F\\
&\quad= n\left[E_F(F_n(x)|Y)-F(x)\right] \quad \text{if } h=1_{(-\infty,x]}\,.
\end{aligned}$$

Therefore the score equations (19) reduce in this case to the equations

$$n\left[E_{\hat F}(F_n(x)|Y)-\hat F(x)\right]=0 \quad \forall x\in\mathbb{R}^d\,, \tag{27}$$

cf. (23). Since $F_n(x)=\frac{1}{n}\sum_i 1_{(-\infty,x]}(X_i)$, it can be verified that when the function g has the special form described above, substituting $F^{(k)}=F^{(k+1)}=\hat F$ in (25) gives exactly (27). Thus a limit of the iterations (25) is a solution of the score equations (19).

Our final example is a simple prototype of the problems which originally motivated this study: Cox's (1972) regression model for which the NPMLE does have all the nice large sample properties one could hope for (see Andersen & Gill 1982; Johansen, 1983; Begun et al, 1983; Dzhaparidze, 1985); and Clayton & Cuzick's (1985a, 1985b) model for dependent survival data, for which very little is known (see Gill 1985; Bickel 1985, 1986). Both these semi-parametric models contain as a special case the non-parametric model of censored survival data with unknown cumulative hazard function. This problem is also a special case of Example 2, with $d=1$ and in which one parametrizes by the function $\Lambda(t)=\int_{[0,t]}(1-F(s-))^{-1}\mathrm{d}F(s)$ instead of by F.

EXAMPLE 3. *Estimation of the cumulative hazard rate with censored data.*
Suppose we have data $(\tilde{X}_i,\Delta_i)$, $i=1,...,n$, where $(\tilde{X}_i,\Delta_i)=(\min(X_i,a_i),\ 1\{X_i\leqslant a_i\})$ for some constants a_i and i.i.d. X_i with d.f. F on $\mathbb{R}_+$ having density (with respect to Lebesgue measure) f, and hazard rate $\lambda=f/(1-F)$. Suppose in fact $a_i\leqslant 1$ for all i so that we can work on the real interval $[0,1]$. The cumulative hazard function Λ is defined (in this case) by

$$\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s \ ;$$

if $F(1)<1$ then $\Lambda(1)<\infty$. In fact $\Lambda(t)=-\log(1-F(t))$ for such continuous F.

We now have a dominated family of distributions of our data, with likelihood function (or Radon-Nikodym derivative)

$$\prod_i f(\tilde{X}_i)^{\Delta_i}\,(1-F(\tilde{X}_i))^{1-\Delta_i} = \prod_i \left[\frac{f(\tilde{X}_i)}{1-F(\tilde{X}_i)}\right]^{\Delta_i}(1-F(\tilde{X}_i))$$

$$= \prod_i \lambda(\tilde{X}_i)^{\Delta_i}\exp(-\Lambda(\tilde{X}_i))\,.$$

Define empirical processes

$$N(t) = \#\{i:\tilde{X}_i\leqslant t,\ \Delta_i=1\}\,,$$

$$Y(t) = \#\{i:\tilde{X}_i\geqslant t\}\ ;$$

observation of these is equivalent to observation of the empirical d.f. of the data. Then we have

$$\log \mathrm{lik} = \int_0^1 \log\lambda(t)N(\mathrm{d}t) - \int_0^1 Y(s)\lambda(s)\mathrm{d}s\,. \tag{28}$$

In fact under many different probability mechanisms for censoring and also under left truncation; see WOODROOFE (1985), ANDERSEN et al (1988); the log likelihood is of precisely this form. (It is also obtained under censored observation of a renewal process). More generally still, we obtain this log likelihood for observation of a *counting process* N in AALEN's (1978) multiplicative intensity model. This model arises in many situations, e.g. in censored observation of time inhomogeneous Markov processes. In some of these models an obvious "discrete" version of the originally "continuous" model does not exist, or several different ones are equally sensible.

We can write the log likelihood ratio of one cumulative hazard function Λ with respect to another, Λ_0, as

$$\int_0^1 \log\left[\frac{\mathrm{d}\Lambda}{\mathrm{d}\Lambda_0}(t)\right]N(\mathrm{d}t) - \int_0^1 Y(s)(\Lambda(\mathrm{d}s)-\Lambda_0(\mathrm{d}s)) \tag{29}$$

(the difference between two versions of (28)). Parametrizing now by Λ instead of by λ, we shall maintain this expression as a log likelihood ratio for *all* finite positive measures Λ, Λ_0 on $[0,1]$ such that $\Lambda \ll \Lambda_0$. In fact this usually only gives the proper answer when Λ and Λ_0 are continuous and the wrong answer when they are discrete; however as far as constructing an estimator and deriving its large sample properties are concerned this should not matter as long as the "true model" has Λ continuous.

Defining $\phi(\Lambda,h,\theta)$ as the cumulative hazard function which is absolutely continuous with respect to Λ with Radon-Nikodym derivative

$$\frac{\mathrm{d}\phi(\Lambda,h,\theta)}{\mathrm{d}\Lambda} = 1+\theta h\,,$$

for $h\in H=\{$ bounded measurable functions on $[0,1]\}$ and θ in some interval around $0\in\mathbb{R}^1$, we can now obtain the likelihood equations

$$\frac{\partial}{\partial\theta}\left[\int_0^1 \log(1+\theta h)\mathrm{d}N - \int_0^1 Y\theta h\,\mathrm{d}\hat{\Lambda}\right]\Big|_{\theta=0}=0$$

for this family: they are simply:

$$\int_0^1 h(\mathrm{d}N - Y\mathrm{d}\hat{\Lambda})=0 \quad \forall h \in H \ ;$$

or equivalently just

$$\int_0^t (\mathrm{d}N - Y\mathrm{d}\hat{\Lambda})=0 \quad \forall t \in [0,1]$$

This has as solution

$$\hat{\Lambda}(t) = \int_0^t \frac{N(\mathrm{d}s)}{Y(s)}, \quad t \in [0,1]$$

which is the well known "empirical cumulative hazard function" or Nelson-Aalen estimator, and which turns up in all the previously mentioned counting process, Markov and semi Markov (Markov renewal) models (see ANDERSEN & BORGAN, 1985, GILL, 1983, ANDERSEN, BORGAN, GILL & KEIDING, 1988, for reviews and further references).

It is especially important to notice in EXAMPLE 3 the appearance of integrals (over an interval in $\mathbb{R}^1$) of one empirical process with respect to another or with respect to the parameter Λ. This is the reason for our detailed look in Section 2.3. at the function $\phi(x,y) \rightarrow \int_{-\infty}^{(\cdot)} x\mathrm{d}y$ mapping $D(\overline{\mathbb{R}})^2$ to $D(\overline{\mathbb{R}})$. The fact that ϕ is not *continuously* differentiable (at least, under the sup norm) rules out (in all interesting examples) the possibility of applying the implicit function theorem when deriving large-sample properties of the solution $\hat{\alpha}$ of (19), considered as a function of a suitably chosen empirical process or distribution function; cf. TSAI & CROWLEY (1985).

Returning briefly to the "extension problem" in EXAMPLE 3, we could also have written the original continuous data likelihood function as

$$\mathrm{lik} = \prod_t \{(\lambda(t)^{\mathrm{d}N(t)}(1 - Y(t)\lambda(t)\mathrm{d}t)^{1-\mathrm{d}N(t)}\}$$

using product integral notation, cf. GILL & JOHANSEN (1987). Thus the log likelihood ratio (29) can also be written as

$$\log\left[\prod_t \left\{ \left[\frac{\mathrm{d}\Lambda}{\mathrm{d}\Lambda_0}(t)\right]^{\mathrm{d}N(t)} \left[\frac{1-Y(t)\mathrm{d}\Lambda(t)}{1-Y(t)\mathrm{d}\Lambda_0(t)}\right]^{1-\mathrm{d}N(t)} \right\}\right] \tag{30}$$

Maintaining this expression for $\Lambda \ll \Lambda_0$ which are not absolutely continuous with respect to Lebesgue measure gives a *different* discrete extension to the model (or rather, its score equations, which is all we are interested in). Coincidentally both (29) and (30) lead to the same NPMLE $\hat{\Lambda}$. However in more complicated versions of these models — Cox's regression model and Clayton & Cuzick's dependent survival times model for instance — the two analogous extensions lead to different NPMLE's. JOHANSEN (1983) essentially chose (29) which is analytically simpler, and that is what counts if one wants simple proofs of large sample properties.

4. ASYMPTOTIC OPTIMALITY OF THE NPMLE

In a forthcoming paper (Part II) it will be shown that if an NPMLE is consistent, *then* it is asymptotically efficient: at least, under a suitable (large) collection of regularity conditions. We restrict attention to the estimation of a cumulative hazard function Λ in an i.i.d. setup modelled after EXAMPLE 3 in Section 3; but this is not the only example covered by any means. One could also add a parametric component so as to cover the Cox regression model or the Clayton & Cuzick dependent survival times (frailty) model. One could also add a parametric component so as to cover the Cox regression model or the Clayton & Cuzick dependent survival times (frailty) model.

One of the regularity conditions will be the assumption that the NPMLE is a Hadamard differentiable function of the empirical d.f. of the data (for the Clayton & Cuzick model this is still an open question). Together with the consistency assumption this forces the functional concerned to yield the true parameter at the true d.f.; and von Mises theory then yields immediately asymptotic normality of $\sqrt{n}(\hat{\Lambda}_n-\Lambda)$. So the main task is to identify the limiting covariance structure and to show that it coincides with the "inverse Fisher information" as generalized to infinite-dimensional parameters by Begun et al. (1983). This is an annoyingly delicate affair; most of the difficulties and new regularity conditions are concerned with our choice of parametrization (Λ itself) and emphasis on *log* likelihood, while the $\mathfrak{L}^2$-based theory of Begun et al. looks at root densities, both of the data and as parametrization (i.e. $\sqrt{d\Lambda/d\Lambda_0}$ for fixed Λ_0 instead of Λ). However the main idea is simple and is modelled on the classical parametric-case proof of asymptotic efficiency of $\sqrt{n}$-consistent solutions-of-likelihood-equations, which goes back to Fisher (1927).

References

1. O. O. Aalen (1978), Nonparametric inference for a family of counting processes, *Ann. Statist.* **6**, 701-726.
2. P. K. Andersen & Ø. Borgan (1985), Counting process models for life history data: a review (with discussion), *Scand. J. Statist.* **12**, 97-158.
3. P. K. Andersen, Ø. Borgan, R. D. Gill, & N. Keiding (1988), *Statistical models for counting processes,* Springer (in preparation).
4. P. K. Andersen & R. D. Gill (1982), Cox's regression model for counting processes: a large sample study, *Ann. Statist.* **10**, 1100-1120.
5. V. I. Averbukh & O. G. Smolyanov (1967), The theory of differentiation in linear topological spaces, *Russian Math. Surveys* **22**, 201-258.
6. V. I. Averbukh & O. G. Smolyanov (1968), The various definitions of the derivative in linear topological spaces, *Russian Math. Surveys* **23**, 67-113.
7. J. M. Begun, W. J. Hall, W.-M. Huang, & J. A. Wellner (1983), Information and asymptotic efficiency in parametric-nonparametric models, *Ann. Statist.* **11**, 432-452.
8. P. J. Bickel (1985), Discussion of papers on semiparametric models at the ISI centenary session in Amsterdam, *Bull. Inst. Int. Statist.* **51**(5), 175-178. (Revised and extended version in: Gill & Voors, eds (1986))
9. P. J. Bickel (1986), *Efficient testing in a class of transformation models,* Technical Report, Berkeley; also in *Papers on semiparametric models at the ISI centenary session, Amsterdam,* see Gill & Voors, eds (1986).
10. P. J. Bickel & D. A. Freedman (1981), Some asymptotic theory for the bootstrap, *Ann. Statist.* **9**, 1196-1217.
11. P. J. Bickel, C. A. J. Klaassen, Y. Ritov, & J. A. Wellner (1987), *Efficient and adaptive estimation in semiparametric models,* John Hopkins University Press, Baltimore.
12. N. E. Breslow & J. Crowley (1974), A large sample study of the life-table and product-limit estimates under random censorship, *Ann. Statist.* **2**, 437-453.
13. M.N. Chang & G. Yang (1988), Strong consistency of a nonparametric estimator of the survival function with doubly censored data, *Ann.Statist.* (to appear).
14. B.R. Clarke (1983), Uniqueness and Frechet differentiability of functional solutions to maximum likelihood type equations, *Ann. Statist.* **11**, 1196-1205.
15. B.R. Clarke (1984), *Nonsmooth analysis and Frechet differentiability of M-functionals,* Preprint #1575, Murdoch Univ.
16. D. Clayton & J. Cuzick (1985a), Multivariate generalizations of the proportional hazards model (with discussion), *J. Roy. Statist. Soc. Ser. A* **148**, 82-117.
17. D. Clayton & J. Cuzick (1985b), The semi-parametric Pareto model for regression analysis of survival times, *Bull. Inst. Int. Statist.* **51**(4), 23.3.1-23.3.18 (for revised and extended version see Gill & Voors, eds (1986)).

18. D. R. Cox (1972), Regression models and life-tables (with discussion), *J. Roy. Statist. Soc. Ser. B* **34**, 187-220.
19. A. P. Dempster, N. M. Laird, & D. B. Rubin (1977), Maximum likelihood estimation for incomplete data via the EM algorithm (with discussion), *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.
20. K. Dzhaparidze (1985), On asymptotic inference about intensity parameters of a counting process, *Bull. Inst. Int. Statist.* **51**(4), 23.2.1-23.2.15 (for revised and extended version see Gill & Voors, eds (1986)).
21. B. Efron (1967), The two sample problem with censored data, *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **4**, 831-883.
22. W. Esty, R. Gillette, M. Hamilton, & D. Taylor (1985), Asymptotic distribution theory for statistical functionals: the compact derivative approach for robust estimators, *Ann. Inst. Statist. Math. Ser. A* **37**, 109-129.
23. L. T. Fernholz (1983), *Von Mises calculus for statistical functionals,* Lecture notes in statistics **19**, Springer Verlag, New York.
24. R. A. Fisher: (1927), On the mathematical foundations of theoretical statistics, *Phil. Trans. Roy. Soc. London* **222**, 309-368; reprinted in R. A. Fisher (1950), *Contributions to mathematical statistics,* 10.309-10.368, Wiley, New York.
25. P. Gaenssler (1984), *Empirical processes,* IMS Lecture notes-monograph series **3**.
26. P. Gaenssler (1987), *Bootstrapping empirical measures indexed by Vapnik-Chervonenkis classes of sets,* Preprint, Univ. of Munich.
27. S. Geman & C.-R. Hwang (1982), Nonparametric maximum likelihood estimation by the method of sieves, *Ann. Statist.* **10**, 401-414.
28. R. D. Gill (1983), Discussion of two papers on dependent central limit theory at the ISI session in Madrid, *Bull. Inst. Int. Statist.* **50**(3), 239-243.
29. R. D. Gill (1985), Discussion of paper by D. Clayton and J. Cuzick, Clayton & Cuzick (1985a) pp. 108-109.
30. R.D. Gill & M.M. Voors, eds (1986), *Papers on semi parametric models at the ISI centenary session* (with discussion), Report MS-R8614, Centrum voor Wiskunde en Informatica, Amsterdam.
31. R. D. Gill & J. A. Wellner (1986), *Large sample theory of empirical distributions in biased sampling models,* Report MS-R8603, Centrum voor Wiskunde en Informatica, Amsterdam (submitted to *Ann. Statist.*
32. R. D. Gill & S. Johansen (1987), *Product integrals and counting processes,* Report MS-R8707, Centrum voor Wiskunde en Informatica, Amsterdam.
33. U. Grenander (1981), *Abstract inference,* Wiley, New York.
34. R. Grü'bel (1988), The length of the shorth, *Ann. Statist.* (to appear).
35. F. R. Hampel, E. M. Ronchetti, P. J. Rousseuw & W. A. Stahel (1986), *Robust Statistics : The Approach Based on Influence Functions,* Wiley, New York.
36. C.C. Heesterman (1987), *A central limit theorem for M-estimators by the von Mises method,* Report MS-R87xx, Centrum voor Wiskunde en Informatica, Amsterdam.
37. M. Jacobsen (1984), Maximum likelihood estimation in the multiplicative intensity model - a survey, *Int. Statist. Rev.* **52**, 193-207.
38. S. Johansen (1978), The product limit estimator as maximum likelihood estimator, *Scand. J. Statist.* **5**, 195-199.
39. S. Johansen (1983), An extension of Cox's regression model, *Int. Statist. Rev.* **51**, 165-174.
40. J. Kiefer & J. Wolfowitz (1956), Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *Ann. Math. Statist.* **27**, 887-906.
41. R.Y. Liu, K. Singh & S.-H. Lo (1986), *On a representation related to the bootstrap,* preprint, Rutgers Univ.
42. K. Lohse (1985), *The consistency of the bootstrap,* Preprint 85-7, Inst f. Math. Stoch., Univ. of Hamburg (to appear in *Statistics & Decisions)*
43. W.C. Parr (1985a), Jackknifing differentiable statistical functionals, *J. Roy. Statist. Soc. Ser.B*

47, 56-66.

44. W.C. Parr (1985b), The bootstrap: some large sample theory and connections with robustness, *Stat. Prob. Letters* **3**, 97-100.
45. D. Pollard (1984), *Convergence of stochastic processes,* Springer, New York.
46. D. Pollard (1985), New ways to prove central limit theorems, *Econometric Theory* **1**, 295-314.
47. J. A. Reeds (1976), *On the definition of von Mises functionals,* Research report S-44, Dept. of statistics, University of Harvard.
48. J.A. Reeds (1978), Jackknifing maximum likelihood estimates, *Ann. Statist.* **6**, 727-739.
49. N. Reid (1981), Influence functions for censored data, *Ann. Statist.* **9**, 78-92.
50. F. W. Scholz (1980), Towards a unified definition of maximum likelihood, *Canad. J. Statist.* **8**, 193-203.
51. R. J. Serfling (1980), *Approximation theorems of mathematical statistics,* Wiley, New York.
52. G. Shorack & J.A. Wellner (1985), *Empirical processes,* Wiley, New York.
53. W. I. Smirnov (1972), *Lehrgang der höheren Mathematik* **5** (4th edition), VEB Deutscher Verlag der Wissenschaften, Berlin (DDR).
54. D. C. Taylor (1985), Asymptotic distribution theory for general statistical functionals, *Ann. Inst. Statist. Math. Ser. A* **37**, 131-138.
55. W.-Y. Tsai & J. Crowley (1985), A large sample study of generalized maximum likelihood estimators from incomplete data via self-consistency, *Ann. Statist.* **13**, 1317-1334
56. B. W. Turnbull (1976), The empirical distribution function with arbitrarily grouped, censored and truncated data, *J. Roy. Statist. Soc. Ser. B* **38**, 290-295.
57. A.W. van der Vaart (1987), *Statistical estimation in large parameter spaces,* Ph.D. Thesis, Univ. of Leiden; to appear in series CWI tracts, Centrum voor Wiskunde en Informatica, Amsterdam.
58. Y. Vardi (1985), Empirical distributions in selection bias models (with discussion by C. L. Mallows), *Ann. Statist.* **13**, 178-205.
59. W. Vervaat (1972), Functional central limit theorems for processes with positive drift and their inverses, *Z. Wahrsch. verw. Gebiete* **23**, 245-253.
60. J. A. Wellner (1985), Semiparametric models: progress and problems, *Bull. Inst. Int. Statist.* **51**(4), 23.1.1-23.1.20 (for revised and extended version see Gill & Voors, eds (1986).
61. M. Woodroofe (1985), Estimating a distribution function with truncated data, *Ann.Statist.* **13**, 163-177.
62. S.-S. Yang (1985), On bootstrapping a class of differentiable statistical functionals with applications to *L*- and *M*-estimators, *Statistica Neerlandica* **39**, 375-385.
63. W.R. van Zwet (1984), A Berry-Esseen bound for symmetric statistics, *Zeit. Wahrsch.verw.Gebiete* **66**, 425-440.