Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

R.D. Gill

Non- and semi-parametric maximum likelihood estimators and the von Mises method (part I)

Department of Mathematical Statistics Report MS-R8604 June

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, January 1986

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-R8604
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 choise 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 suprising 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.[1]

Let B_1 and B_2 be two locally convex topological vector spaces. In the sequel these spaces will often be normed and complete, i.e. Banach spaces, but unfortunately will usually not be separable. Let $\phi: B_1 \to B_2$ 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 $d\phi(x): B_1 \to B_2$ we define the remainder of ϕ at $x+h$, $\mathrm{Rem}(x+h)$, by

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

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 $d\phi(x).h = 0$ when $h=0$, and $\mathrm{Rem}(x+0)=0$ too. We will say that ϕ is differentiable at x, with derivative $d\phi(x)$ at that point, if $\mathrm{Rem}(x+h)$ is of smaller order than h as h tends to 0:

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

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

1. See also Parr (1985a,b).

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

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

($S \subseteq B_1$ is called bounded if for any neighbourhood U of $0 \in B_1$, $\lambda U \supseteq S$ for all sufficiently large $\lambda \in \mathbb{R}^+$.) Clearly bounded differentiability (of ϕ at x) implies compact differentiabillity, and that implies directional differentiability. The derivative $d\phi(x)$ remains the same. In applications one often determines the form of the derivative by computing the Gateaux derivative acting on h, $d\phi(x).h$, 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 en 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 $d\phi(x)$ 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. These limitations of Gateaux and Frechet differentiation are well illustrated by the results in SERFLING (1980; chapter 6).[1] On the other hand Hadamard differentiability is exactly attuned to statistical applictions and nicely separates analytical considerations about ϕ from probabilistic considerations about F_n. Consider the following theorem, in which X_n might play the role of an empirical distribution function, considered as a random element of some topological vector space, and μ would then be the true distribution function:

THEOREM 1. (THE δ-METHOD.) *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*

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

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 the usual 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

1. But see also B. R. CLARKE (1983), Uniqueness and Frechet differentiability of functional solutions to maximum likelihood type equations, *Ann. Statis.* **11**, 1196-1205.

differentiability, drawing the following correspondences:

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

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

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

For reasons we come to later (measurability problems !) we shall hardly ever use exactly THEOREM 1, but for the time being it should motivate our further exploration of Hadamard differentiability. Also 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.

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 (when B_1 is a metric space)

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

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 $n\to\infty$.

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

$$\mathrm{Rem}(x+th_n)\to 0, \quad \forall\ t\to 0, \quad \forall\ h_n\to h\in H\,. \tag{6}$$

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_1 is a Banach 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 $d\psi(\phi(x)).d\phi(x)$ (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 $d\phi(x).d\psi(0)$.

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 $d_1\psi$ and $d_2\psi$ of ψ with respect to x and y : One expects

$$d\phi(x) = -[d_2\psi(x, \phi(x))]^{-1} d_1\psi(x, \phi(x)) .$$

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 $d\psi(x,y)$ varies continuously (with respect to the topology of uniform convergence on compact subsets of $B_1 \times B_2$; see REEDS (1976) Appendix A)[1] 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.[2] 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.

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 $\phi=\psi^{-1}: B_1 \to B_2$ of a given mapping $\psi : B_2 \to B_1$.

1. With *continuous* differentiability, Gateaux, Hadamard end Frechet theories more or less coincide; see REEDS (1976; Appendix A). For bootstrap and jackknife applications it seems as though continuous differentiability is required and hence in effect Frechet differentiability; see REEDS (1978), BICKEL & FREEDMAN (1981) and PARR (1985a,b). However an important role is still played by choice of topology.
2. This is a very delicate matter. There are errors is REID (1981) and in CROWLEY & TSAI (1985) concerning exactly this point.

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 1, to proving asymptotic normality of a statistical quantity, cons idered as a function of the empirical distribution function. Some of the problems are due to our naieve choice of topology on the space of distribution functions: namely the topology based on the supremum norm.

The examples concern independent and identically distributed observations on the real line. For the sake of familiarity we assume in fact that the observations are in the interval [0,1]; however this is just a question of notation. The restriction to the real line is crucial, and it is perhaps only for distribution functions on $\mathbb{R}^1$ that the supremum norm is an appropriate metric at all. 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 [0,1] respectively, we look at asymptotic normality of $\phi(F_n)=F_n^{-1}(p)$, $p\in(0,1)$ and of $\phi(F_n,G_m)=\int_0^1 F_n(x)\mathrm{d}G_m(x)$. We want to obtain these results by using only the well-known weak convergence of $n^{\frac{1}{2}}(F_n-F)$ in $D[0,1]$ (and similarly for G_m) and differentiability of the function ϕ in each case. The first example is purely illustrative; however the second is relevant to non-parametric maximum likelihood estimation since the functional $(F,G)\to\int F\mathrm{d}G$ plays an important role in very many of the examples from survival analysis, Markov processes, etc.

To start with we consider a general one-sample functional $\phi(F_n)$ with F_n considered as an element of $D[0,1]$. It is well-known that $\sqrt{n}(F_n-F)$ converges weakly to the process $B^0\circ F$, where B^0 is a Brownian bridge on [0,1], and weak convergence holds with respect to the Skorohod topology on $D[0,1]$ defined by one of the Skorohod metrics J. Unfortunately $(D[0,1],J)$ is not a topological vector space: addition is not a continuous operation with this topology. So an immediate application of Theorem 1 is thwarted! We therefore work with $(D[0,1],\|\cdot\|)$ where $\|\cdot\|$ denotes the supremum norm on $D[0,1]$: $\|x\|=\sup_{t\in[0,1]}|x(t)|$. Now $D[0,1]$ is a topological vector space, in fact it is a Banach space, but non-separable. We must now investigate weak convergence and tightness of $n^{\frac{1}{2}}(F_n-F)$ in this new space. However an immediate problem is that $n^{\frac{1}{2}}(F_n-F)$ is not even a random element of $(D[0,1],\|\cdot\|)$: i.e. the mapping from the underlying probability space Ω (on which the random sample of size n is defined) to $D[0,1]$ is generally not measurable. Consider the case $n=1$ and $F=$ uniform distribution on [0,1]. So $F_1(t)=1_{[U,1]}(t)$ where $U\tilde{}$ Uniform [0,1]. Let $B\subset[0,1]$ be arbitrary, and let

$$\mathcal{N}_t=\{x\in[0,1]:\|x-1_{[t,1]}\|<\tfrac{1}{2}\}\ ,$$

Now $\mathcal{N}_t$ is open so $\cup_{t\in B}\mathcal{N}_t=O_B$ is open too. But the subsets of Ω $\{U\in B\}$ and $\{F_1\in O_B\}$ are identical. So if the mapping from $\omega\in\Omega$ to $F_1\in D[0,1]$ were measurable, $\{U\in B\}$ would be an event for every $B\subset[0,1]$. In particular it would be possible to assign a probability to $\{U\in B\}$ for *every* $B\subset[0,1]$. But this is equivalent to extending Lebesgue-measure $\sigma-$additively to all subsets of [0,1], which is impossible.

This embarrassing but only technical problem can be avoided in several ways. Reeds takes an approach based on property (5) of Hadamard differentiation and inner probability arguments. He shows that for continuous F, $\forall\epsilon>0$, $\exists\delta_n\to0$ and a compact $K\subset(D[0,1],\|\cdot\|)=B_1$ such that

$$P_*(\mathrm{dist}(\sqrt{n}(F_n-F),K)\leqslant\delta_n)\geqslant 1-\epsilon\quad\forall n$$

(P_* denotes inner probability). Then (5) shows that if $\phi:B_1\to B_2$ is Hadamard differentiable at F, and if

$$\sqrt{n}\mathrm{Rem}(F_n)=\sqrt{n}(\phi(F_n)-\phi(F))-\mathrm{d}\phi(F).\sqrt{n}(F_n-F)$$

is a random variable (!) then $\sqrt{n}\mathrm{Rem}(F_n) \xrightarrow{P} 0$ as $n \to \infty$. In each specific application easy direct arguments show that $d\phi(F).\sqrt{n}(F_n - F)$ is asymptotically normal ($d\phi(F)$ is a linear map so the central limit theorem can be applied) giving at last the required result.

Here we use an alternative solution (mentioned by Reeds too) based on property (4) of Hadamard differentiation and the Skorohod-Dudley theorem. According to this theorem if X_n is a sequence of random elements of a metric space such that $X_n \xrightarrow{\mathcal{D}} X$ as $n \to \infty$, then there exists a probability space with defined on it $X'_n \overset{\mathcal{D}}{=} X_n$ and $X' \overset{\mathcal{D}}{=} X$ such that $X'_n \xrightarrow{a.s.} X'$. This is known as a Skorohod-Dudley construction. We also need the specific fact that if $x_n \to x$ in $(D[0,1],J)$ and x is continuous then $x_n \to x$ in $(D[0,1], \|\cdot\|)$. These ingredients allow us to prove the following theorem, which replaces Theorem 1 in applying Hadamard differentiation to our examples:

THEOREM 2. *Suppose $\phi:(D[0,1], \|\cdot\|) \to \mathbb{R}$ 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 $X_1, \ldots, X_n$ from a continuous distribution F on $[0,1]$. Then*

$$n^{\frac{1}{2}}(\phi(F_n) - \phi(F)) \xrightarrow{\mathcal{D}} d\phi(F).Z \tag{7}$$

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

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

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

$$\mathrm{I}\, C(\phi; F, x) = \lim_{t \to 0} \frac{\phi((1-t)F + t1_{[x,1]}) - \phi(F)}{t}\,.$$

We sketch the main part of the proof of this theorem, ignoring measurability questions. Since $\sqrt{n}(F_n - F) \xrightarrow{\mathcal{D}} Z$ in $(D[0,1],J)$ we are guaranteed the existence of a probability space with, defined on it, random elements $Z'_n \overset{\mathcal{D}}{=} \sqrt{n}(F_n - F)$, $Z' \overset{\mathcal{D}}{=} Z$ and $Z'_n \xrightarrow{a.s.} Z$ (J). From Z'_n we can recover random elements $F'_n = n^{-\frac{1}{2}} Z'_n + F$; this is an empirical distribution function and $F'_n \overset{\mathcal{D}}{=} F_n$ (J). Since Z' has continuous sample paths when F is continuous, we have $Z'_n \xrightarrow{a.s.} Z'$ $(\|\cdot\|)$. Now Hadamard differentiability and (4) give immediately

$$\sqrt{n}(\phi(F'_n) - \phi(F)) \xrightarrow{a.s.} d\phi(F).Z'\,.$$

By equality of distributions we conclude (7).[1]

A few remarks on this theorem are in order. Firstly, note that in the proof we only actually needed that ϕ is differentiable tangentially to $C[0,1]$; cf. definition (6). This fact is vital for our applications. Secondly, the theorem was stated and proved for the empirical distribution function in the i.i.d. case and for real-valued ϕ, so that we could add the characterization of the limiting distribution in terms of the well-known influence curve, but the general idea of using a Skorohod-Dudley construction in order to switch from convergence (J) to convergence $(\|\cdot\|)$ has far wider applicability. Finally, returning to the special i.i.d. case, we could also have given a version of Theorem 2 in which F is not

1. For a recent survey of the influence curve, see HAMPEL et al. (1986).

required to be continuous, since a Skorohod-Dudley construction with sup-norm convergence is still possible by means of some extra tricks. Then of course we do need proper Hadamard differentiability of ϕ.

Now that we have a usable theorem we can turn to our specific examples. They illustrate a new collection of problems: how to extend a given ϕ from { d.f.'s on $[0,1]$} to all of $D[0,1]$. They also demonstrate the usefulness of the concept of differentiability tangentially to a subspace. Reeds, Fernholz and more lately ESTY et al. (1985) and TAYLOR (1985) treat this and similar problems by constructing continuous modifications of empirical distribution functions and then working in $C[0,1]$, avoiding both measurability and differentiability difficulties. On the other hand 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.

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

$$F_n(T_n-)\leqslant p\leqslant F_n(T_n) \tag{8}$$

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 in the set of distribution functions, and (8) holds. Now we claim that we can extend ϕ to all x in a neighbourhood in $(D[0,1], \|\cdot\|)$ of the true distribution function F in such a way that the analogue of (8) still holds:

$$x(\phi(x)-)\leqslant p\leqslant x(\phi(x)) \tag{9}$$

for all x. For instance, for each $x\in D[0,1]$ which is not a distribution function, *define* $\phi(x)=\sup\{t:x(t)\leqslant p\}$. Now we show that such a function is Hadamard differentiable tangentially to $C[0,1]$ at a point $x=F$ which is a distibution function, differentiable at its p'th quantile with a positive derivative there. To make the notation lighter we shift p and the p'th quantile to the origin and work with $D[-1,1]$ instead of with $D[0,1]$.

LEMMA 1 *Let $x\in D[-1,1]$ be fixed and nondecreasing, and satisfy $x(0)=0$, x is differentiable at 0 with positive derivation $x'(0)$. Let h_n be a sequence of elements of $D[-1,1]$ and t_n a sequence of elements of $\mathbb{R}^+$ such that that $h_n\overset{\|\cdot\|}{\to}h\in C[-1,1]$ and $t_n\to 0$ as $n\to\infty$. Define $x_n=x+t_nh_n$ and suppose $\theta_n\in[-1,1]$ satisfies*

$$x_n(\theta_n-)\leqslant 0\leqslant x_n(\theta_n)\quad \forall n\ . \tag{10}$$

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 $1/t_n$.

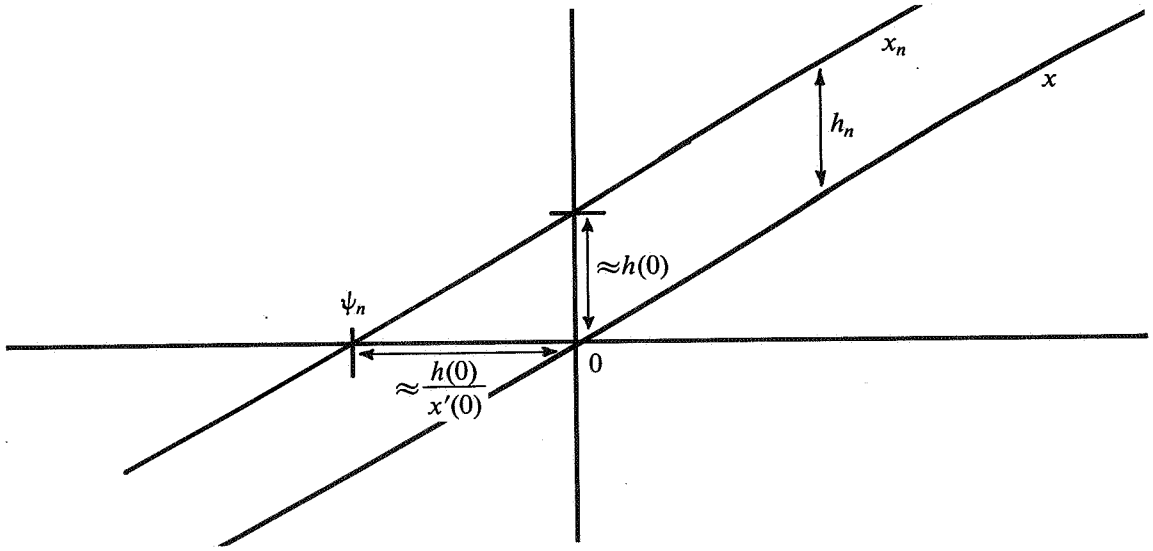

FIGURE 1 : Derivative of $\phi(x)=x^{-1}(0)$

PROOF OF LEMMA 1. 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 $u\in[-M,M]$; so $\psi_n\in[-M,M]$ too for all large enough n. We have for $u\in[-M,M]$

$$x_n(t_nu) = x(t_nu)+t_nh_n(t_nu) \tag{11}$$

$$x(t_nu) = t_nu.x'(0)+o(t_n) \text{ uniformly in } u\,. \tag{12}$$

So substituting (12) in (11) and (11) in (10) with $u=\psi_n$, $\theta_n=t_n\psi_n$, we obtain

$$t_n\psi_n\,.x'(0)+o(t_n)+t_nh_n(t_n\psi_n-) \leqslant 0 \leqslant \tag{13}$$

$$\leqslant t_n\psi_n.x'(0)+o(t_n)+t_nh_n(t_n\psi_n)$$

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 (13) throughout by t_n (lim sup ψ_n).$x'(0)+h(0)\leqslant 0\leqslant$(lim inf ψ_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

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

$$x(u)\geqslant ca \quad a\leqslant u\leqslant 1$$

Let $A<\infty$ be an upper bound to $|h_n|$ on $[-1,1]$ for all n. Then

$$x_n(u) = x(u)+t_nh_n(u)\geqslant\begin{cases} cu-t_nA & 0\leqslant u\leqslant a \\ ca-t_nA & a\leqslant u\leqslant 1\end{cases}$$

Thus if n is sufficiently large that $ca-t_nA>0$, we have $x_n(u)>0$ for $u>t_nA/c$ (see Figure 2). 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 $|\psi_n|\leqslant A/c<\infty$. □

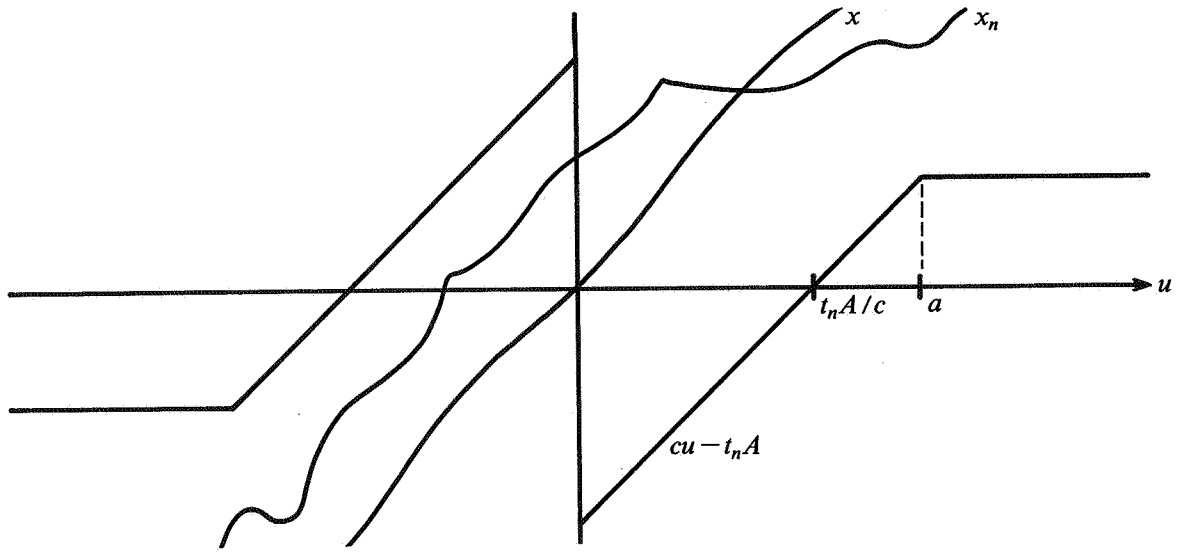

FIGURE 2 : Proof of lim sup$|\psi_n|<\infty$.

Taking the remarks after (4) and (5) in subsection 2.2. into account, we obtain the following corollary:

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

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

for all y in some neighbourhood of x. Then ϕ is Hadamard differentiable at x tangentially to $C[0,1]$ with derivative

$$d\phi(x).h \;=\; -h(\theta)/x'(\theta)\,.$$

Note that the derivative is indeed a continuous linear map from $(D[0,1],\|\cdot\|)$ to $\mathbb{R}$. Combining this with Theorem 2 gives :

COROLLARY 2. *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 continuous distribution F with $F(\theta)=p$, F differentiable at θ with $F'(\theta)>0$, then*

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

We have actually shown that ϕ is Hadamard differentiable tangentially to $\{h\in D[0,1]:h$ is continuous at $\theta\}$. This enables one with a little extra effort to drop the restriction in Theorem 2 that F is continuous (though the conditions on F at θ are still needed !).

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 continuous sample paths of the limiting process. Also the method can be extended to give, via differentiability of a suitably defined extension of the

mapping $F \to F^{-1}$, weak convergence of the whole quantile process. For a similar approach see VERVAAT (1972).

In our second example we are again confronted with the problem of extending a functional of empirical distribution functions to $D[0,1]$. We omit the details of the application to the Wilcoxon statistic, but just recall that this can be contructed from the mapping $(F,G) \to \int_{-\infty}^{\infty} F\mathrm{d}G$ for two distribution functions F and G. We shall investigate the differentiability of this rather simple mapping. Surprisingly this is not a trivial matter.

Consider a mapping $\phi:(D[0,1])^2 \to \mathbb{R}$ which is such that $\phi(x,y) = \int_0^1 x\mathrm{d}y$ for $x,y \in D[0,1]$ which are both non-decreasing. How can we define $\phi(x,y)$ for other x and y, which may for instance both be of unbounded variation ? One possible definition is

$$\phi(x,y) = \int_0^1 x\mathrm{d}y^{\uparrow}$$

where $y^{\uparrow}$ is the smallest non-decreasing function which is larger than or equal to y itself. We shall later show this (arbitrary) choice of extension is not crucial; for the time being we just take this particular choice for the sake of convenience.

LEMMA 2. *The mapping $\phi:(x,y) \to \int_0^1 x\mathrm{d}y^{\uparrow}$ from $(D[0,1])^2$ to $\mathbb{R}$ is Hadamard differentiable with respect to the supremum norm tangentially to $D[0,1]\times C[0,1]$ at a point (x,y) which is such that x is non-decreasing and y is strictly increasing. The derivative is given by*

$$\mathrm{d}\phi(x,y).(h,k) = \int_0^1 x\mathrm{d}k + \int_0^1 h\mathrm{d}y$$

where the first integral (with respect to k) is defined by formal integration by parts.

Before proving the Lemma we note that the form of the derivative is easily established by computing the partial derivatives of ϕ with respect to x and y separately. Also we have not specified whether the integration from 0 to 1 is over the interval $(0,1]$ or $[0,1]$ (in the latter case one usually adds the convention $y(0-)=0$.) The result is true in both cases as long as the same convention is used throughout.

PROOF. Suppose the sequences $t_n \in \mathbb{R}^+$, $h_n \in D[0,1]$ and $k_n \in D[0,1]$ satisfy $t_n \to 0$, $h_n \to h \in D[0,1]$ and $k_n \to k \in C[0,1]$. Let (x,y) be as described. We must establish that

$$\begin{aligned} & t_n^{-1}\ \mathrm{Rem}(x+t_nh_n, y+t_nk_n) \qquad (14)\\ & = t_n^{-1}\Big(\int (x+t_n\,h_n)\mathrm{d}(y+t_nk_n)^{\uparrow} - \int x\mathrm{d}y - t_n\int x\mathrm{d}k_n \\ & \qquad - t_n\int h_n\mathrm{d}y\Big) \to 0 \ \text{ as } \ n\to\infty\ , \end{aligned}$$

where the integrals are taken over $(0,1]$ or $[0,1]$ as appropriate. We must also verify that

$$\Big(\int x\mathrm{d}k_n\ , \int h_n\mathrm{d}y\Big) \to \Big(\int x\mathrm{d}k\ , \int h\mathrm{d}y\Big) \ \text{ as } \ n\to\infty\ ,$$

i.e. that the derivative is a continuous linear map, but this is easy. Now the expression on the right hand side of (14) can be rewritten as the sum of two terms:

$$\begin{aligned} & t_n^{-1}\ \Big(\int x\mathrm{d}(y+t_nk_n)^{\uparrow} - \int x\mathrm{d}y^{\uparrow} - t_n\int x\mathrm{d}k_n\Big) \qquad (15)\\ & \quad + \Big(\int h_n\mathrm{d}(y+t_nk_n)^{\uparrow} - \int h_n\mathrm{d}y^{\uparrow}\Big)\ . \end{aligned}$$

To show the first term converges to zero, by integrating by parts it suffices to show that

$$t_n^{-1}((y+t_nk_n)^{\uparrow} - y^{\uparrow}) - k_n \to 0 \ \text{ as } \ n\to\infty \qquad (16)$$

i.e. differentiability of the mapping "$\uparrow$", tangentially to $C[0,1]$, at a point y which is strictly increasing. For the second term of (15), it suffices to show that

$$\int h_n dy_n \to \int h dy \tag{17}$$

if $(h_n,y_n)\to(h,y)$ in $(D[0,1],\|\cdot\|)$ and satisfies furthermore lim sup$\int|dy_n|<\infty$, and hence also $\int|dy|<$lim sup$\int|dy_n|<\infty$.

To prove (16) we note that $k_n\to k$, with $k\in C[0,1]$, implies that

$$\forall\epsilon>0\ \ \exists n_0,\delta: n\geqslant n_0\ ,|u-v|<\delta\Rightarrow|k_n(u)-k_n(v)|<\epsilon\ .$$

Suppose also $\|k_n\|\leqslant c<\infty\forall n$ and

$$\inf(y(u+\delta)-y(u))=\eta>0\ \ 0\leqslant u\leqslant 1-\delta$$

(here we use the assumption that y is strictly increasing). Now for $t_n<\eta/2c$, $n\geqslant n_0$ we find for each $v\in[0,1]$

$$\begin{aligned} y(v)+\ t_nk_n(v)&\leqslant\sup_{u\leqslant v}\ (y(u)+t_nk_n(u))\ = \\ &=\sup_{v-\delta\leqslant u\leqslant v}(y(u)+t_nk_n(u))\leqslant y(v)+t_nk_n(v)+t_n\epsilon \end{aligned} \tag{18}$$

because for $u<v-\delta$, we have $y(u)<y(v)-\eta$ and $t_nk_n(u)<t_nk_n(v)+2t_nc$, hence

$$\begin{aligned} y(u)+t_nk_n(u)\ &<y(v)+t_nk_n(v)+2t_nc-\eta< \\ &<y(v)+t_nk_n(v)\ . \end{aligned}$$

But from (18) we find

$$|t_n^{-1}\left[(y+t_nk_n)^{\uparrow}(v)-y(v)\right]-k_n(v)|\leqslant\epsilon$$

Since ϵ was arbitrary, this establishes (16).

To prove (17) we note that for given $\epsilon>0$ one can choose $0=u_0<u_1<\cdots<u_m=1$ such that

$$v\in[u_i,u_{i+1})=I_i\Rightarrow\begin{cases}|h_n(v)-h_n(u_i)|<\epsilon\ \ \forall n\ , \\ |h(v)-h(u_i)|<\epsilon\end{cases}.$$

Now

$$\begin{aligned} &\int_0^1 h_n dy_n-\int_0^1 h dy=\sum_i(\int_{I_i}h_n dy_n-\int_{I_i}h dy)= \\ &=\sum_i\left[h_n(u_i)\int_{I_i}dy_n-h(u_i)\int_{I_i}dy\right] \\ &+\sum_i\left[\int_{I_i}(h_n-h_n(u_i))dy_n-\int_{I_i}(h-h(u_i))dy\right]\ . \end{aligned}$$

As $n\to\infty$ the first sum converges to zero. The second sum, in absolute value, is less than or equal to $\epsilon(\int_0^1|dy_n|+\int_0^1|dy|)$. So

$$\text{lim sup}\ |\int_o^1 h_n dy_n-\int_0^1 h dy|\leqslant 2\ \epsilon\ \text{lim sup}\int_0^1|dy_n|$$

Since ϵ was arbitrary, this gives the required result. □

Careful inspection of this proof shows that one can actually prove differentiability tangentially to $D[0,1]\times\{k\in D[0,1]$: k is continuous where y is continuous, constant on intervals where y is constant $\}$ at a point (x,y) where x and y are both non-decreasing. Thus we can obtain a perfectly general result on asymptotic normality of the Wilcoxon two-sample test statistic (i.e. without any continuity or strict monotonicity restrictions on the underlying distribution functions F and G). Also, the result can easily be extended to prove differentiability tangentially to $D[0,1]\times C[0,1]$ of the mapping

$\phi:(D[0,1])^2 \to D[0,1]$ defined by $\phi(x,y) = \int_0^{\cdot} x\,dy^{\uparrow}$. This can be applied in the many examples from survival analysis (e.g. estimation of a cumulative hazard rate, k-sample tests) which involve the functional $\phi:(x,y) \to \int_0^{\cdot} x\,dy^{\uparrow}$.

However there is also a negative aspect to LEMMA 2. The functional ϕ is only differentiable at a point in $(D[0,1])^2$ satisfying monotonicity properties, and is clearly not differentiable in a whole *neighbourhood* of such a point, and certainly not continuously differentiable.[1] 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 $D[0,1]$.

The proof of Lemma 2 appears quite complicated, and one may wonder whether or not a simpler proof is possible. In fact the last part of the proof is actually a standard result from analysis called Helly's theorem (see 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.[2] 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.

We close this section with a discussion of the role of the particular extension (from distribution functions to $D[0,1]$) which has to be chosen for each functional before differentiability can be verified. The following Lemma shows that this choice is irrelevant. So our problem with verifying differentiability in the examples did not derive from an inappropriate extension.

LEMMA 3. *Suppose $x \in E \subset B_1$, $\phi:E \to B_2$, and $\overline{E}$ is a neighbourhood of x. Suppose there exists a continuous linear map $d\phi(x):B_1 \to B_2$ such that for all $t_n \to 0$ ($t_n \in \mathbb{R}$) and $h_n \to h \in B_1$ such that $x_n = x + t_n h_n \in E$ for all n,*

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

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

PROOF. The existence of an extension, differentiable at x, is easily established by the choice $\phi(x+h) = \phi(x) + d\phi(x).h$ 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 $s + t'h' \in E$ shows that this extension is indeed differentiable with derivative $d\phi(x)$. 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. For such functionals as $(x,y) \to \int_0^{\cdot} x\,dy$, naturally defined for x and y of bounded variation on $[0,1]$, one can note that the set of functions of bounded variation in $D[0,1]$ is dense in $D[0,1]$ (under the supremum norm topology).

1. One can easily exhibit sequences $x_n \to x$, $k_n \to k$ such that $\int x_n\,dk_n \not\to \int x\,dk$.
2. One can complete a von Mises treatment of the product-limit estimator by proving Hadamard differentiability of the functional $x \to \prod_0^{\cdot}(1 + dx)$; see JOHANSEN (1977), GILL & JOHANSEN (1987).

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) and Johansen (1983).[1] 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 $\{P_\alpha : \alpha \in \mathcal{A}\}$, where $\mathcal{A}$ may be infinite-dimensional and P_α is the distribution of X on the sample space $\mathcal{X}$ under α, if

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

for all $\alpha \in \mathcal{A}$ and $\mu \gg P_{\hat{\alpha}}, P_\alpha$; so we take a different μ – e.g. $P_\alpha + P_{\alpha'}$ – when comparing each $\alpha, \alpha' \in \mathcal{A}$ (Scholz addresses the problem that $dP_\alpha/d\mu$ is only defined a.e. $-\mu$, 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 $\{P_\alpha : \alpha \in \mathcal{A}\}$ where the parameter space $\mathcal{A}$ 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 $\phi(\alpha, h, \theta) \in \mathcal{A} : \alpha \in \mathcal{A}, h \in H, \theta \in \mathbb{R}$ such that $\phi(\alpha, h, 0) = \alpha$ for all h. Thus for each $\alpha \in \mathcal{A}$ 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

1. Also Jacobsen (1984) and Wang (1986). 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.

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:

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

where

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

and

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

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 $\alpha\in\mathcal{A}$ to $\alpha\in\bar{\mathcal{A}}$ for some larger set $\bar{\mathcal{A}}$, 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 $\mathcal{A}$ 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 distibution function $\phi(F,h,\theta)$ absolutely continuous with respect to F by

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

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

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

So

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

and

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

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 $(X_i 1_{B_i}(X_i), 1_B(X_i))$ 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. 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).

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 X has density $f_Y(y;\theta)$ too. Usually we will then have

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

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

$$n\left\{E_{\hat{F}}(F_n(x)|Y)-\hat{F}(x)\right\}=0 \quad \forall x\in\mathbb{R}^d\,, \tag{26}$$

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 (26). 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 almost nothing is known (see GILL 1985; BICKEL 1985).[1] 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 $(\tilde{X}_i,\Delta_i)$, $i=1,\ldots,n$, 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 Lebesque 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)

$$\begin{aligned}
\prod_i f(\tilde{X}_i)^{\Delta_i}\,(1-F(\tilde{X}_i))^{1-\Delta_i} &= \prod_i\left\{\frac{f(\tilde{X}_i)}{1-F(\tilde{X}_i)}\right\}^{\Delta_i}(1-F(\tilde{X}_i)) \\
&= \prod_i \lambda(\tilde{X}_i)^{\Delta_i}\exp(-\Lambda(\tilde{X}_i))\,.
\end{aligned}$$

1. But see also BICKEL (1986).

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

$$\log \text{lik} = \int_0^1 \log\lambda(t)N(\mathrm{d}t) - \int_0^1 Y(s)\lambda(s)\mathrm{d}s\ . \tag{27}$$

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

$$\int_0^1 \log\left[\frac{\mathrm{d}\Lambda}{\mathrm{d}\Lambda_0}(t)\right] N(\mathrm{d}t) - \int_0^1 Y(s)(\Lambda(\mathrm{d}s) - \Lambda_0(\mathrm{d}s)) \tag{28}$$

(the difference between two versions of (27)). 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

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

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 $\phi(x,y) \to \int_0^{\cdot} x \mathrm{d}y^{\uparrow}$ mapping $D[0,1]^2$ to $D[0,1]$. 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 also could also have written the original continuous data likelihood function as

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

using product integral notation, cf. JOHANSEN (1977), GILL & JOHANSEN (1987). Thus the log likelihood ratio (28) can also be written as

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

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 (28) and (29) 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 analoguous extensions lead to different NPMLE's. JOHANSEN (1983) essentially choose (28) which is analytically simpler, and that is what counts if one wants simple proofs of large sample properties.

4. ASYMPTOTIC OPTIMALITY OF THE NPMLE

(to be continued in Part II).

In this section 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 convered by any means.[1] 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. 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 $\mathcal{L}^2$-based theory of Begun et al. looks at root densities, both of the data and as parametrization (i.e. $\sqrt{\mathrm{d}\Lambda/\mathrm{d}\Lambda_0}$ 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).

1. One could also add a parametric component so as to cover the Cox regression model or the Clayton & Cuzick dependent survival times (frailty) model.

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 (to appear).
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. Suveys* **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. Int. Statist. Inst.* **51**(4), 23.4.1-23.4.4. (Revised and extended version in: *Papers on semiparametric models at the ISI centenary session (with discussion),* R. D. GILL & M. M. VOORS (eds.), Report MS-R86xx, Centrum voor Wiskunde en Informatica, Amsterdam).
9. P. J. BICKEL (1986), *Efficient testing in a class of tranformation models,* Technical Report, Berkeley; also in *Papers on semiparametric models at the ISI centenary session, Amsterdam,* see BICKEL (1985).
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. D. CLAYTON & J. CUZICK (1985a), Multivariate generalizations of the proportional hazards model (with discussion), *J. Roy. Statist. Soc. Ser. A* **148**, 82-117.
14. D. CLAYTON & J. CUZICK (1985b), The semi-parametric Pareto model for regression analysis of survival times, *Bull. Int. Statist. Inst.* **51**(4), 23.3.1-23.3.18 (for revised and extended version see BICKEL, 1985).
15. D. R. COX (1972), Regression models and life-tables (with discussion), *J. Roy. Statist. Soc. Ser. B* **34**, 187-220.
16. 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.
17. K. DZHAPARIDZE (1985), On asymptotic inference about intensity parameters of a counting process, *Bull. Int. Statist. Inst.* **51**(4), 23.2.1-23.2.15 (for revised and extended version see BICKEL, 1985).
18. B. EFRON (1967), The two sample problem with censored data, *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **4**, 831-883.
19. 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.
20. L. T. FERNHOLZ (1983), *Von Mises calculus for statistical functionals,* Lecture notes in statistics **19**, Springer Verlag, New York.
21. 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.

22. S. GEMAN AND C.-R. HWANG (1982), Nonparametric maximum likelihood estimation by the method of sieves, *Ann. Statist.* **10**, 401-414.
23. R. D. GILL (1983), Discussion of two papers on dependent central limit theory at the ISI session in Madrid, *Bull. Int. Statist. Inst.* **50**(3), 239-243.
24. R. D. GILL (1985), Discussion of paper by D. Clayton and J. Cuzick, *CLAYTON & CUZICK (1985a)* pp. 108-109.
25. 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.*).
26. R. D. GILL & S. JOHANSEN (1987), *Product integrals and counting processes* (in preparation).
27. U. GRENANDER (1981), *Abstract inference,* Wiley, New York.
28. F. R. HAMPEL, E. M. RONCHETTI, P. J. ROUSSEUW & W. A. STAHEL (1986), *Robust Statistics : The Approach Based on Influence Functions,* Wiley, New York.
29. M. JACOBSEN (1984), Maximum likelihood estimation in the multiplicative intensity model - a survey, *Int. Statist. Rev.* **52**, 193-207.
30. S. JOHANSEN (1977), *Product integrals and Markov processes,* Preprint 3, Inst. of Math. Statist., University of Copenhagen.
31. 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.
32. W.C. PARR (1985a), Jackknifing differentiable statistical functionals, *J. Roy. Statist. Soc. (B)* **47**, 56-66.
33. W.C. PARR (1985b), The bootstrap: some large sample theory and connections with robustness, *Stat. Prob. Letters* **3**, 97-100.
34. J. A. REEDS III (1976), *On the definition of von Mises functionals,* Research report S-44, Dept. of statistics, University of Harvard.
35. REEDS (1978), Jackknifing maximum likelihood estimates, *Ann. Statist.* **6**, 727-739.
36. N. REID (1981), Influence functions for censored data, *Ann. Statist.* **9**, 78-92.
37. F. W. SCHOLZ (1980), Towards a unified definition of maximum likelihood, *Canad. J. Statist.* **8**, 193-203.
38. R. J. SERFLING (1980), *Approximation theorems of mathematical statistics,* Wiley, New York.
39. W. I. SMIRNOV (1972), *Lehrgang der höheren Mathematik* **5** (4th edition), VEB Deutscher Verlag der Wissenschaften, Berlin (DDR).
40. D. C. TAYLOR (1985), Asymptotic distribution theory for general statistical functionals, *Ann. Inst. Statist. Math. Ser. A* **37**, 131-138.
41. 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 (see also correction note, *Ann. Statist.* **14**, xx-xx).
42. B. W. TURNBULL (1976), The empirical distribution function with arbitrarily grouped, censored and truncated data, *J. Roy. Statist. Soc. Ser. B* **38**, 290-295.
43. Y. VARDI (1985), Empirical distributions in selection bias models (with discussion by C. L. Mallows), *Ann. Statist.* **13**, 178-205.
44. W. VERVAAT (1972), Functional central limit theorems for processes with positive drift and their inverses, *Z. Wahrsch. verw. Geb.* **23**, 245-253.
45. M.-C. WANG (1986), *Product-limit estimates: a generalized maximum likelihood study,* Preprint, Dept. of Biostatistics, Johns Hopkins University.
46. J. A. WELLNER (1985), Semiparametric models: progress and problems, *Bull. Int. Statist. Inst.* **51**(4), 23.1.1-23.1.20 (for revised and extended version see *BICKEL, 1985*).
47. M. WOODROOFE (1985), Estimating a distribution function with truncated data, *Ann. Statist.* **13**, 163-177.