Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

R.D. Gill, M.N. Voors (eds.)

Papers on semiparametric models at the ISI centenary session, Amsterdam

Department of Mathematical Statistics Report MS-R8614 November

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.).

Copyright © Stichting Mathematisch Centrum, Amsterdam

Papers on Semiparametric Models

at the ISI Centenary Session, Amsterdam

R.D. Gill, M.N. Voors (eds.)
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

Report MS-R8614
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

Bibliotheek
Centrum voor Wiskunde en Informatica
Amsterdam

Forword

Richard D. Gill
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

At the ISI Centenary Session in Amsterdam, August 1985, I had the great pleasure of organizing a morning's meeting on semiparametric models. This report contains revised and extended versions of the papers presented at that meeting together with the invited and open discussion.[1] As a sequel we have included an extra paper by Peter Bickel which complements his discussion.

It seems to me that the area of semiparametric models is a very exciting one right now, as is atestified by Jon Wellner's fine survey paper which opens this collection. It lies at the centre of developments in 'classical' mathematical statistics: extending asymptotic estimation theory from the finite dimensional to the infinite dimensional case, and synthesizing parametric and nonparametric approaches to statistics. At the same time it is being fuelled by developments in applied statistics. The typical semiparametric models which any self-respecting theory must be able to cope with have arisen in practical statistics, especially in the vigorous field of survival analysis; and the prototype of them all is Cox's celebrated regression model. Survival analysis continues to provide tougher and tougher models for the theory to cope with, as David Clayton and Jack Cuzick's paper and Peter Bickel's discussion and extra paper show.[2]

Via survival analysis and especially the Cox regression model there is a very strong link, represented here by Kacha Dzhaparidze's paper, to the statistical analysis of stochastic processes and especially the Scandinavian and Russian approaches based on the 'French' (no longer just 'Strasbourg') school of stochastic analysis. So with a strong international flavour we have both applied and theoretical, probabilistic and statistical, ingredients.

One cannot make general remarks on developments in statistics nowadays without mentioning the influence of the computer. This is self-evident as far as the theory and practice of 'infinite-dimensional' statistical models are concerned. Future developments, and not just practical ones, will surely have a strongly computational flavour. However the mere physical existence of the present report was inspired by the possibilities of computer word-processing and electronic mail, which will have just as large an influence on research in statistics. Each paper was prepared at the author's home institute and on arrival in Amsterdam by various electronic means was reformatted and typeset on the CWI's UNIX system according to our local conventions. This turned out to be a remarkably painless and amusing operation, despite such complications as the fact that one of the papers started

1. The original versions appear in *Bull. Int. Statist. Inst.* **51** (4,5), Meeting 23 (Amsterdam, 1985).
2. For my own opinion of how one should treat this model, see my discussion of Clayton & Cuzick's (1985) *JRSS (A)* paper, and CWI report MS-R8604 on NPMLE in general.

life on a rather British personal computer. Here I must especially thank my co-editor Michel Voors, who took care of most of the dirty work, aided by programmer Rob van der Horst.

Finally some more words of thanks. We are especially indebted to the ISI for their kind permission to produce and distribute this reprint, and last but not least very grateful to the speakers and discussants at the meeting for their enthusiastic support of this project. We hope their enthusiasm for the field of semiparametric models will be equally infectious.

Amsterdam, July 31, 1986.

Table of Contents

Semiparametric Models: Progress and Problems

Jon A. Wellner
Department of Statistics
B313 Padelford Hall
University of Washington
Seattle, Washington 98195
U.S.A.

Semiparametric models, models which incorporate both parametric (finite-dimensional) and nonparametric (infinite-dimensional) components, have received increasing use and attention in statistics in recent years. This paper reviews developments in this very large and rich class of models which spans the middle ground between parametric and nonparametric models. Attention is devoted to a preliminary classification of such models with comments on recent work, to lower bounds for estimation, to two potentially useful methods for construction of efficient estimates, and to open problems.

1. Introduction

Models for phenomena involving randomness play a key role in statistics. If $\mathbf{P}_{all}$ denotes the collection of all probability distributions on a sample space $\mathbf{X}$ of the observations X, a model $\mathbf{P}$ is a subset of $\mathbf{P}_{all}$: thus we assume in constructing a model $\mathbf{P}$ that X has a distribution P in $\mathbf{P}$, and we write $X \cong P \in \mathbf{P}$. The sample space $\mathbf{X}$ is the set of all possible observations.

A statistician uses the observations X to make inferences about the 'true' probability distribution P, and hence about real-world phenomena in question. A common form of inference is *point estimation.* For example, if X represents the life expectancy or survival time of an individual who has been given a new medical treatment, the statistician may be interested in using a sample of such individuals to estimate $\nu(P) \equiv P(X > t)$, the probability of survival beyond t time units. The choice of a model $\mathbf{P}$ can have a major effect on inferences about $\nu(P)$: If the model $\mathbf{P}$ is too small, the statistician runs the risk that the model will not contain the 'true' P, and the consequent price is bias in estimation of $\nu(P)$. In this case the model is not sufficiently large to be realistic and may fail to capture the essential features of the phenomena in question. On the other hand, if the model $\mathbf{P}$ is too large, the statistician may find himself in the position of estimating too many parameters from too little data. This tradeoff between realism and parsimony is an ever-present theme in statistics; for interesting discussions of some aspects of model-building see Chapters 2 and 4 of Cox and Snell [23] or Stone [76].

Parametric models $\mathbf{P}_0 \equiv \{P_\theta : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^d$ for some d play a dominant role in classical statistical theory. Such models, with a finite-dimensional parameter space Θ, form the basis of much of classical statistics. For example, life expectancies or survival times are frequently modeled by the parametric family of exponential distributions with $P_\theta(X > t) = \exp(-\theta t)$, $t \geq 0$, $\theta > 0$. In this case inferences about P_θ can essentially be reduced to inference about θ. Once an estimator $\hat{\theta}$ of θ is available, then other functions $\nu(P_\theta) = q(\theta)$ of P can be estimated as $q(\hat{\theta}) = \nu(P_{\hat{\theta}})$. A difficulty with such parametric models is that typically a parametric model $\mathbf{P}_0$ is a relatively small subset of $\mathbf{P}_{all}$, and hence the 'true' distribution P of X may not be contained in $\mathbf{P}_0$.

One approach to this difficulty is the completely nonparametric approach: assume only that $P \in \mathbf{P}_{all}$ or a slight restriction of $\mathbf{P}_{all}$ requiring only some smoothness or monotonicity assumptions. In this case of life expectancies or survival times, all probability distributions P on the sample space $X = [0, \infty)$ would be considered and $P = P_{all}$.While this approach seems to be feasible when the dimensionality of the sample space is small, it fails to take advantage of structure in the phenomena being modeled and begins to run into difficulty when the dimensionality of the sample space (and hence of the parameter space, $\mathbf{P}_{all}$ itself) is large.

A compromise strategy which gains in model realism and the flexibility needed to make use of the larger data sets which are increasingly available is the semiparametric approach: assume that some aspects or components of the model are parametric or finite-dimensional, while other aspects or components are allowed to be nonparametric or infinite-dimensional. Then the resulting *semiparametric model* $\mathbf{P}$ is typically of the form

$$\mathbf{P} = \{P_{\theta,G}: \ \theta \in \Theta, \ G \in \mathbf{G}\}$$

where $\Theta \subset \mathbb{R}^d$ for some d and $\mathbf{G}$ is some (large) collection of functions. We also write

$$\mathbf{P} = \{P_\theta: \ \theta = (\theta_1, \theta_2) \text{ with } \theta_1 \in \Theta_1 \subset \mathbb{R}^d, \ \theta_2 \in \Theta_2\},$$

where Θ_2 is a collection of functions.

This semiparametric approach has proved to be very useful in a wide range of problems, and promises to play an increasingly important role in statistics. Our object here is to survey this extremely rich and flexible class of models (Section 2), and to briefly review the developing inference methods with emphasis on lower bounds for estimation and construction of efficient estimates of the parametric component of such models (Section 3 and 4). The survey of models and review of inference methods may be read independently of one another. The final section discusses open problems.

The notion of a semiparametric model is very general, and is already being used, at least implicitly, in situations involving observations which are not independent and identically distributed (iid). For simplicity, however, we restrict attention here to the iid case: throughout this paper $X_1, \ldots, X_n$ are iid according to the distribution $P \in \mathbf{P}$ where $\mathbf{P}$ is a parametric or semiparametric model.

2. Classes of semiparametric models

Little effort has been made to classify or categorize semiparametric models. While such an effort may be premature, it may also help to identify related models and aid in developing methods to apply to new problems. The following scheme should be regarded as provisional and temporary.

The classification of models given here has two fundamental categories: *basic models*, and *derived models*. The basic models consist of exponential family models, group models, and transformation models. The derived models include regression models, convolution models, mixing models, censoring models, and biased sampling models. Although this scheme is both redundant and possibly incomplete, it includes all the semiparametric models with which I am now familiar. The rest of this section elaborates on these categories, and provides examples of the models of the various types with some brief comments on recent work.

2.1. Basic Models

The following basic models serve as building blocks in the construction of semiparametric models.

2.1.1. Exponential family models. (A). These are familiar parametric models with density (with respect to some measure m)

$$p(x,\theta) = c(\theta)\exp(\sum_{i=1}^{k} Q_i(\theta)T_i(x))h(x)$$

for $\theta\in\Theta\subset\mathbb{R}^k$, $x\in\mathbf{X}\subset\mathbb{R}^d$. While these are themselves completely parametric (finitely dimensional) models, they serve as building blocks for many interesting semiparametric models.

2.1.2. Group models. (B).

(1). The classical parametric model of this type is obtained as follows: suppose that $Y\cong G\equiv P_0$, a fixed distribution on $\mathbf{X}$, and let $\mathbf{V}$ denote a group of (one to one) transformations on $\mathbf{X}$ parametrized by $\theta\in\Theta\subset\mathbb{R}^k$. If $v_\theta\in\mathbf{V}$, let $X\equiv v_\theta(Y)\cong P_\theta$ for $\theta\in\Theta$.

Examples:

(a) Location. $\mathbf{X}=\mathbb{R}^d$, $v_\theta(x)=x+\theta$ with $\theta\in\mathbb{R}^d$, and $P_\theta=P_0(\cdot-\theta)$.

(b) Elliptic distributions. $\mathbf{X}=\mathbb{R}^d$, $v_\theta(x)=\theta^{-1/2}x$ where θ is positive definite and symmetric; $G\equiv P_0$ is spherically symmetric on $\mathbb{R}^d$. Then $\mathbf{P}=\{P_\theta:\theta\in\Theta\}$ is the P_0-family of elliptic distributions.

(c) Two-sample models. $\mathbf{X}=\mathbf{X}_0\times\mathbf{X}_0$, $\mathbf{V}=\mathbf{V}_0\times\mathbf{V}_0$ where $\mathbf{V}_0$ is a group of transformations on $\mathbf{X}_0$, $\theta=(\mu,\nu)\in\Theta_0\times\Theta_0\equiv\Theta$, $Y=(W,Z)$ with $W,Z\cong P_0$ independent, and $X=(v_\mu(W),v_\nu\circ v_\mu(Z))$.

(2). By letting the distribution P_0 in (1) range over some large class of probability distributions $\mathbf{G}$ small enough to still allow identification of θ, or at least some important functions of θ, yields a semiparametric model

$$\mathbf{P}=\{P_{\theta,G}:\theta\in\Theta,G\in\mathbf{G}\}.$$

Examples:

(a) If $\mathbf{X}=\mathbb{R}^1$ in 1(a) above and $\mathbf{G}$ is the family of distributions symmetric about 0, $\mathbf{P}$ is the classical symmetric location family.

(b) If $\mathbf{X}$ and Θ are as in 1(b) above and $\mathbf{G}$ is the family of all spherical symmetric distributions, then $\mathbf{P}$ is the family of all elliptic distributions; see e.g. BICKEL [6].

(c) If $\mathbf{X}$ and Θ are as in 1(c) and $\mathbf{G}$ is arbitrary, then ν is still identifiable; see STEIN [73] or PFANZAGL [66].

(3). Classical nonparametric statistical theory uses transformation groups which are not parametrizable by a Euclidean space; for example, all continuous monotone transformations from $\mathbb{R}$ to $\mathbb{R}$. See LEHMANN [51] page 24 and 25 for 'semiparametric subgroups' of the large group and note that examples 2(a) and 2(b) are of this type. A wealth of other 'semiparametric group' families are undoubtedly possible.

2.1.3 Transformation models. (C). These models typically map $(\theta,P)\to P_\theta$ where $\theta\in\Theta\subset\mathbb{R}^k$ and $P\in\mathbf{G}$, a collection of probability distributions on $\mathbf{X}$. The key feature is that the map $P_\theta=\psi(\theta,P)$ acts on P, or some function that is one-to-one with P, rather than on X as in the case of a group model.

The classical example of this type of model is that of a family of 'Lehmann alternatives' defined as follows (see LEHMANN [50]): Let $\mathbf{X}=\mathbb{R}^1$, suppose that $Y\cong G$ and let $\{B(\cdot,\theta):\theta\in\Theta\subset\mathbb{R}^k\}$ be a family of monotone transformations from $[0,1]$ to $[0,1]$ with $B(0,\theta)=0$, $B(1,\theta)=1$ for all $\theta\in\Theta$. Then $X\cong P_{\theta,G}$ has df (distribution function) $F_{\theta,G}(x)=B(G(x),\theta)$. Here are some particular cases.

Examples:

(a) $B_a(\mu,\theta)=1-(1-\mu)^\theta$ with $0<\theta<\infty$. This yields the *proportional hazards model*: $\Lambda_F(x)=\theta\Lambda_G(x)$ where Λ_F is the cumulative hazard function corresponding to F; see LEHMANN [50] and COX [22].

(b) $B_b(\mu,\theta)=\dfrac{\theta\mu}{\theta\mu+(1-\mu)}=\dfrac{\theta\mu(1-\mu)^{-1}}{1+\theta\mu(1-\mu)^{-1}}$ with $0<\theta<\infty$. This yields the *proportional odds model*

$$\frac{F(x)}{1-F(x)}=\theta\frac{G(x)}{1-G(x)};$$

see BENNETT [2].

(c) $B_c(\mu,\theta,\nu)=1-[1-\nu\theta\log(1-\mu)]^{-1/\nu}$, $0<\nu<\infty$, $\theta>0$. This yields the *semiparametric Pareto model* suggested by CLAYTON AND CUZICK [19]. Note that $B_c(\mu,\theta,\nu)\to B_a(\mu,\theta)$ as $\nu\to 0$ while Bennett's B_b is related to Clayton and Cuzick's B_c by

$$B_c(1-\exp(-\frac{\mu}{1-\mu}),\theta,1)=B_b(\mu,\theta).$$

These three models can all be written in the form

$$h(X)=-\log(\theta)+\epsilon \tag{2.1}$$

where $h(x)\equiv\log\Lambda_G(x)=\log[-\log(1-G(x))]$ and ϵ has the distribution:
(a) $F(x)=1-\exp(-e^x)$ (extreme value);
(b) $F(x)=1/(1+e^{-x})$ (logistic);
(c) $F(x)=1-1/(1+\nu x)^{1/\nu}$ (Pareto).

Because of the generality allowed for the transformations h, rank methods and partial likelihoods play an important role in analyzing these models. Note that (1) yields a transformation family linear model if $\theta=\exp(\gamma z)$, and shows that these models can be viewed as special cases of a type of model involving smooth transformations of both X and z considered by BREIMAN AND FRIEDMAN [11]; see 2.2.1 below and DOKSUM [24].

2.2 Derived models

The following classes of models are all derived from the basic models given above.

2.2.1 Regression models. (D). Given a basic model of one of the three types described above, there is a straightforward recipe for constructing related regression models:

1. Start with an exponential family, group or transformation model $\mathbf{P}=\{P_{\theta,G}:\theta\in\Theta,G\in\mathbf{G}\}$ where θ is the finite-dimensional Euclidean component of the model and G is the nonparametric or infinite-dimensional component of the basic model.
2. Suppose that $Z\cong H$ on $\mathbb{R}^d$.
3. Given $Z=z$, replace θ (or a component thereof) in the basic model by a semiparametric regression function $r(\gamma,z)$ taking values in Θ where $\gamma\in\Gamma\subset\mathbb{R}^k, k>0$. Different forms for r ranging from parametric to nonparametric regression models, with many interesting intermediate semiparametric forms, are possible. For example:
   (a) Linear model: $r(\gamma,z)=\gamma z$;
   (a′) Exponential linear model: $r(\gamma,z)=\exp(\gamma z)$;
   (b) Nonlinear: $r(\gamma,z)=r_0(\gamma,z)$ for a fixed known nonlinear function r_0;
   (c) Nonparametric: $r(\gamma,z)=r(z)$, with r smooth (see HUBER [37]);
   (d) Semiparametric: $r(\gamma,z)=\gamma z_1+r(z_2)$, where $z=(z_1,z_2)$, and r is smooth;
   (e) Projection pursuit: $r(\gamma,z)=r(\gamma z)$ where $|\gamma|=1$ and $r:\mathbb{R}^1\to\mathbb{R}^1$ is smooth;
   (f) Signal-noise: $r(\gamma z)$ where $r:\mathbb{R}^1\to\mathbb{R}^1$ is periodic with period 1 so that γ is a frequency parameter.

Combining various types of regression functions illustrated by (a) - (f) with the basic models A, B or C yields a rich collection of regression models, including parametric, semiparametric, and

nonparametric models. STONE [76] gives an interesting survey and further references.
Examples (with brief comments concerning recent work):

(a) Combining basic model A with the regression model D(a) yields linear exponential family regression models; see e.g. LEHMANN [51] Chapter 3, pages 196 - 207.

(b) Combining the basic model B1(a) where P_0 is normal with D(a) yields classical parametric normal theory regression models; the extension to B2(a) yields semiparametric linear regression models with arbitrary (symmetric) error distributions.

(c) The basic model B1(a) (with P_0 a fixed distribution on $\mathbb{R}^1$; e.g. normal) combined with the semiparametric regression model D(d) leads to a very interesting class of regression models introduced by ENGLE, GRANGER, RICE AND WEISS [26] to study effects of weather on electricity demand, and by WAHBA [81]. This model has one nonparametric component, the smooth regression function r. Generalizations with two nonparametric components by allowing the error distribution to be arbitrary are also of interest. A special case has been studied by SCHICK [72], while STONE [76] discusses a spectrum of related regression models.

(d) Combining B2(a) with D(e) leads to a model related to projection-pursuit regression; see FRIEDMAN AND STUETZLE [27], STONE [76], and HUBER [37].

(e) Combining C(a) with D(a′) yields Cox's [22] proportional hazards model. Many variants on this model are possible and deserve further exploration. Replacement of the exponential with some other (fixed) non-negative function has been considered by PRENTICE AND SELF [69], while C(c) combined with D(a′) has been explored by CLAYTON AND CUZICK [19]. TIBSHIRANI [78] considers a version of Cox's model with the linear function in $\exp(\gamma z)$ replaced by a sum of smooth but otherwise arbitrary functions $\sum_{i=1}^{k} r_i(z_i)$. See 2.2.2 below for related mixture models involving unobserved covariates.

(f) Combination of B1(a) or B2(a) with D(f) yields a semiparametric 'signal plus noise' model which extends classical parametric signal plus noise models. For the latter, see IBRAGIMOV AND HAS'MINSKII [38]. MCDONALD [58] has some interesting preliminary work on semiparametric extensions. These models are of interest in astrophysical applications; see e.g. LAFLER AND KINMAN [44] or STELLINGWERF [74].

2.2.2 Mixture models. (E). Mixture models can usually be viewed as the result of unobserved heterogeneity as follows: suppose that $X=(Y,Z)$ has a distribution of the form

$$P_{\theta,G,H}(Y\in A, Z\in B) = \int_B P_{\theta,G}(Y\in A \mid Z = z)dH(z).$$

Then if we can only observe Y, the observations have the *mixture distribution*

$$P_{\theta,G,H}(Y\in A) = \int P_{\theta,G}(Y\in A \mid Z = z)dH(z).$$

Examples:

(a) Paired exponentials. Suppose that $(Y\equiv(Y_1,Y_2)\mid Z=z)\cong$ (exponential (z),exponential(θz)):

$$f(y \mid z) = \theta z^2 \exp(-(zy_1 + \theta z y_2))1_{[0,\infty)}(y_1)1_{[0,\infty)}(y_2)$$

and suppose $Z\cong H$ on $\mathbb{R}^+$. Then

$$f(y)\equiv f_{\theta,H}(y) = \int_0^\infty \theta z^2 \exp(-z(y_1+\theta y_2))dH(z);$$

see e.g. LINDSAY [53]. Here θ is a parametric component and H a nonparametric component of the model, and the mixed distribution is parametric while the mixing distribution is nonparametric. Generalizations of this model, including regression type models, have been studied and advocated for use in modeling micro-economic data by HECKMAN AND SINGER [35].

(b) Dependent proportional hazards or frailty models. Suppose that $(Y\equiv(Y_1,Y_2)\mid Z=z)$ has joint

survival function

$$P_G(Y_1 \geqslant y_1, Y_2 \geqslant y_2 \mid Z = z) = [1 - G_1(y_1)]^z [1 - G_2(y_2)]^z$$

with $G = (G_1, G_2)$ and suppose that $Z \cong \text{Gamma}(\nu, \lambda)$. Then with $\theta = (\nu, \lambda)$,

$$P_{\theta,G}(Y_1 \geqslant y_1, Y_2 \geqslant y_2) = \frac{\lambda^\nu}{[\lambda + \Lambda_1(y_1) + \Lambda_2(y_2)]^\nu}$$

where $\Lambda_i \equiv -\log(1 - G_i)$, $i = 1,2$. In this case the mixed distribution is nonparametric while the mixing distribution is a parametric family. This model, which serves as an alternative to (a), has been studied by CLAYTON [16] and OAKES [65], and has been generalized by GILL [28], who derives nonparametric maximum likelihood estimators, see also the discussion to [17]. Related regression models are discussed by RIDDER AND VERBAKEL [70] and ELBERS AND RIDDER [25].

(c) Errors in variables models. Suppose that $X = (Y, Z)$ with

$$Y_1 = Z + \epsilon_1$$

$$Y_2 = \alpha + \beta Z + \epsilon_2$$

where $Z \cong H$ (non-Gaussian) and $\epsilon \equiv (\epsilon_1, \epsilon_2) \cong N(0, \Sigma)$. The resulting mixture model is an *errors in variables regression* model. Consistent maximum likelihood estimates were obtained by KIEFER AND WOLFOWITZ [42], but lower bounds for estimation of (α, β) together with asymptotically efficient estimates attaining the bounds were first obtained by BICKEL AND RITOV [9].

(d) If $(Y \mid Z = z) \cong$ exponential (z) and $Z \cong H$, then

$$P_H(Y \geqslant y) = \int_0^\infty \exp(-yz) dH(z).$$

Estimation of H via nonparametric maximum likelihood methods in this and more general situations has been considered by LAIRD [45] and JEWELL [39]. While the estimates are known to be consistent, little is known about the efficiency of the estimates or their rate of convergence.

Other results concerning mixing models and efficient estimation have also been obtained by LAMBERT AND TIERNEY [46], [47], and by HAS'MINSKII AND IBRAGIMOV [34].

2.2.3 Censoring models. (F). These models are derived from other models of one of the above types as follows: Suppose that $X \cong P_{\theta,G} \in \mathbf{P}$, and suppose that T is a many-to-one function on the sample space $\mathbf{X}$ of X. Then we can observe only $X^* \equiv T(X) \cong P^*_{\theta,G}$.

Examples:

(a) Mixing. The mixing models of E are censoring models with $X^* \equiv T(Y,Z) = Y$.

(b) Random right censorship. In this type of censoring, which has received much use in survival analysis, $X^* \equiv (X_1^*, X_2^*) \equiv T(X_1, X_2) \equiv (X_1 \wedge X_2, 1_{[X_1 \leqslant X_2]})$. Random right censoring meshes extremely well with Cox's proportional hazards regression model as discussed in D(e). On the other hand, however, this type of censoring can make estimation quite difficult. For example, estimation for the linear regression model D(b) with arbitrary right censoring of the dependent variable has been considered by MILLER [63] and by BUCKLEY AND JAMES [13]; see also HALPERN AND MILLER [62]. RITOV [71] has, in spite of the difficulties, computed information lower bounds and produced asymptotically efficient estimators achieving the bounds. TIBSHIRANI [77] considered a version of this censored regression model with the linear (parametric) regression function replaced by a smooth regression function.

(c) Convolution. Here $X^* \equiv T(X_1, X_2) \equiv X_1 + X_2$ where X_1 and X_2 are independent. The traffic model of BRANSTON [10] is a model which results from this convolution type of censoring combined with a simple mixture model. Further results on this model are given by P. Groeneboom

and A. Koning in CWI reports MS-N8401 and MS-R8508 respectively.

2.2.4 Biased sampling models. (G). Suppose that $X \cong P_{\theta,G} \in \mathbf{P}$, a semiparametric model. Then suppose that $K_i(x)$, $i=1,\ldots,s$ is a collection of known non-negative *biasing kernels* and that λ_i, $i=1,\ldots,s$ is a probability distribution on $\{1,\ldots,s\}$. Then the *biased sampling distribution* corresponding to $P_{\theta,G}$, $\underline{K}=(K_1,\ldots,K_s)$, and $\underline{\lambda}=(\lambda_1,\ldots,\lambda_s)$ is

$$P_{\theta,G,\lambda}(X \in A, I=i) = \frac{\int_A K_i(x) P_{\theta,G}(dx)}{\int_{\mathbf{X}} K_i(x) P_{\theta,G}(dx)} \lambda_i \tag{2.2}$$

for $i=1,\ldots,s$. Here are some examples of this type of model.

Examples:

(a) Vardi's selection bias model. Suppose that $P_{\theta,G}=G$ and $K_1,\ldots,K_s$ are biasing functions with $\int K_i dG < \infty$ for $i=1,\ldots,s$, and $\lambda_i \geqslant 0$ satisfy $\sum_{i=1}^s \lambda_i = 1$. Then

$$P_{G,\lambda}(X \in A, I=i) = \frac{\int_A K_i dG}{\int_{\mathbf{X}} K_i dG} \lambda_i, \quad i=1,\ldots,s.$$

VARDI [80] gives a condition which guarantees existence of the nonparametric maximum likelihood estimate of G. The particular case with $\mathbf{X}=\mathbb{R}^1$, $K_1(x)=1, K_2(x)=x$, which involves the length-biased distribution $\int_0^x y dG(y)/\mu$ corresponding to G was studied by VARDI [79], and the further special case with $\lambda_1=0=1-\lambda_2$ was considered earlier by COX [21]. Consistency, asymptotic normality, and efficiency of Vardi's nonparametric maximum likelihood estimator are addressed in a forthcoming paper by GILL AND WELLNER [29].

(b) Choice-based sampling models. Suppose that $X \equiv (Y,Z)$, where $Z \cong H$ is a vector of covariates, and $(Y \mid Z=z) \cong \text{Multinomial}_k(1, p(\theta,z))$ (where k denotes the number of cells and the number of trials is 1); we will write $[Y=y]$ for the event that outcome y occurs, $y=1,\ldots,k$. A frequently used model for the p's is the multinomial - logit model with

$$P_\theta(Y=y \mid Z=z) = p_y(\theta,z) = \frac{\exp(\theta_y z)}{\sum_{y'=1}^k \exp(\theta_{y'} z)},$$

but in any case this part of the model is parametric; the nonparametric part of the model is G. To get a 'choice-based sampling model', let $K_i(x) \equiv K_i(y,z) = 1_{D_i}(y)$, $i=1,\ldots,s$ where $D_1,\ldots,D_s$ are known subsets of $\{1,\ldots,k\}$. Then the biased sampling model (2) becomes

$$P_{\theta,G}(Y=y, Z \in B, I=i) = \frac{\int_B 1_{D_i}(y) P_\theta(Y=y \mid Z=z) dG(z)}{\int \sum_{y=1}^k 1_{D_i}(y) P_\theta(Y=y \mid Z=z) dG(z)} \lambda_i.$$

This type of model has received considerable use in biostatistics and econometrics; see PRENTICE AND PIKE [68], BRESLOW [12], COSSLETT [20] for some history and further references. Estimation for this model was considered by MANSKI AND LERMAN [56]. The efficiency of their estimators of θ and generalizations were treated by COSSLETT [20]. In general the 'choice functions' or biasing kernels may depend on both y and z; see MANSKI AND MCFADDEN [57]

(c) Truncated regression models. Suppose that $X=(Y,Z)$ with $Y=\theta Z+\epsilon$ where $\epsilon \cong G$ with density g and $Z \cong H$ are independent. Thus the basic semiparametric model is a linear regression model with unknown error distribution G. If $s=1$ and $K(x)=K(y,z)=1_{(-\infty,y_0]}(y)$ where y_0 is a fixed

constant, then

$$P_{\theta,G}(Y\in A, Z\in B) = \frac{\int_B \int_{(-\infty,y_0]\cap A} g(y-\theta z)dydH(z)}{\int\int_{(-\infty,y_0]} g(y-\theta z)dydH(z)}.$$

This truncated regression model has been investigated by BHATTACHARYA, CHERNOFF AND YANG [5]. Motivated by a controversy in astronomy concerning Hubble's law, they constructed $\sqrt{n}$-consistent estimators of the regression parameters θ. Further results for this model have been obtained by JEWELL [41], who also gives additional examples. JEWELL [40] has also considered estimation for generalizations of this model with $s\geqslant 2$ corresponding to stratified sampling on the dependent variable Y.

3. BOUNDS FOR ESTIMATION

Lower bounds for the variances of estimators play an important role in statistical theory, setting a baseline or standard against which estimators can be compared. In their classical form such bounds assert that *any* unbiased estimator $\hat{\theta}_n$ of $\theta, \theta\in\mathbb{R}_1$, has variance no smaller than $(nI(\theta))^{-1}\equiv b(\theta)/n$:

$$\mathrm{Var}_\theta[\hat{\theta}_n]\geqslant \frac{b(\theta)}{n}$$

In other words $b(\theta)/n$ is the smallest variance we can hope for in an unbiased estimator $\hat{\theta}_n$ of θ. If $\hat{\theta}_n^b$ is an estimator which asymptotically *achieves* the bound (in the sense that $\sqrt{n}(\hat{\theta}_n^b-\theta)\to_d N(0,b(\theta))$), then we say that $\hat{\theta}_n^b$ is *asymptotically efficient.* If the statistician uses an estimator $\hat{\theta}_n^a$ which is inefficient, then he has not used the data to best advantage and is essentially wasting observations. Hence if $\hat{\theta}_n^a$ is another estimator with $\sqrt{n}(\hat{\theta}_n^a-\theta)\to_d N(0,a(\theta))$ where $a(\theta)\geqslant b(\theta)$ necessarily, then the limiting ratio of sample sizes which yields equal standard deviations (and hence also equal variances) of $\hat{\theta}_n^b$ and $\hat{\theta}_n^a$ is called the *asymptotic relative efficiency* $e_{a,b}$ of $\hat{\theta}_n^a$ with respect to $\hat{\theta}_n^b$; evidently $e_{a,b}=b(\theta)/a(\theta)\leqslant 1$. If the estimator $\hat{\theta}_n^a$ has asymptotic relative efficiency 1/2 relative to an (efficient) estimator $\hat{\theta}_n^b$ and the estimator $\hat{\theta}_n^b$ requires $n_b=100$ observations to yield a given variance, then $n_a=200$ observations will be needed to achieve the same variance using the inefficient estimator $\hat{\theta}_n^a$; half the data are 'wasted' by the use of $\hat{\theta}_n^a$. Thus in the search for 'good' estimators and other inference procedures, statisticians are interested in answers to the questions: A. How well can we do? What are the lower bounds for estimation in the model at hand? B. How can we construct efficient estimates, i.e. estimates which achieve the bounds?

Our aim in this section is to briefly survey classical (Cramér - Rao) and modern (Hajek - Le Cam) bounds for estimation in 'regular' parametric models. The Hajek - Le Cam approach has led to the development of lower bounds for estimation in nonparametric and semiparametric models. Bounds of this type have been established by BERAN [3], KOSHEVNIK AND LEVIT [43], LEVIT [52], MILLAR [59], [60], [61], PFANZAGL [66], and BEGUN et al. [1]. We give a brief introduction to these bounds for semiparametric models at the end of this section. A thorough treatment will be given in the forthcoming monograph by BICKEL, KLAASSEN, RITOV AND WELLNER [7].

3.1. Cramér - Rao lower bounds

First consider the case of a 'regular' parametric model: suppose that $X_1,\ldots,X_n$ are iid $P_\theta\in\mathbf{P}\equiv\{P_\theta:\theta\in\Theta\}$ where $\Theta\subset\mathbb{R}^d$ is open, that $\mathbf{P}$ is dominated by a (sigma-finite) measure μ on $\mathbf{X}$, and let $p(\cdot,\theta)\equiv\frac{dP_\theta}{d\mu}$ for $\theta\in\Theta$. Then the classical *log-likelihood* of an observation X is

$$l(\theta,X)\equiv\log p(X,\theta),$$

the *scores vector* $\dot{l}$ is

$$\dot{\mathbb{l}}(\theta,X) \equiv \nabla \mathbb{l}(\theta,X) = \frac{1}{p(X,\theta)}\left(\frac{\partial}{\partial\theta_1}p(X,\theta), \ldots, \frac{\partial}{\partial\theta_d}p(X,\theta)\right)^{\mathsf{T}},$$

and the *Fisher information matrix* for θ is

$$I(\theta) = E_\theta[\dot{\mathbb{l}}(\theta,X)\dot{\mathbb{l}}(\theta,X)^{\mathsf{T}}].$$

Assume that $I(\theta)$ is positive definite so that $I(\theta)^{-1}$ exists.

One form of the classical Crámer-Rao inequality for unbiased estimates $a^{\mathsf{T}}\hat{\theta}_n$ of $a^{\mathsf{T}}\theta$, where a is a fixed vector in $\mathbb{R}^d$, is:

$$n\mathrm{Var}_\theta[a^{\mathsf{T}}\hat{\theta}_n] \geqslant a^{\mathsf{T}}I(\theta)^{-1}a = \sup_{b\in\mathbb{R}^d}\frac{(a^{\mathsf{T}}b)^2}{b^{\mathsf{T}}I(\theta)b}. \tag{3.1}$$

If we focus on estimation of the first component $\theta_1 \in \mathbb{R}^1$ of θ, it follows immediately from (1), the definition of $I(\theta)$, and standard L_2-projection or regression theory that

$$\begin{aligned} n\mathrm{Var}_\theta[\hat{\theta}_1] &\geqslant \sup_{b\in\mathbb{R}^d}\frac{b_1^2}{b^{\mathsf{T}}I(\theta)b} \equiv I^{11}(\theta) \qquad (3.2)\\ &= \frac{1}{\inf_{c\in\mathbb{R}^d, c_1=1} E_\theta[\dot{\mathbb{l}}_1 - c_2\dot{\mathbb{l}}_2 - \ldots - c_d\dot{\mathbb{l}}_d]^2}\\ &= \frac{1}{I_{11}(\theta) - I_{12}(\theta)I_{22}^{-1}(\theta)I_{21}(\theta)} \equiv \frac{1}{I_{11}^*(\theta)} \end{aligned}$$

where

$$I(\theta) \equiv \begin{bmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{bmatrix}, \quad I(\theta)^{-1} = \begin{bmatrix} I^{11}(\theta) & I^{12}(\theta) \\ I^{21}(\theta) & I^{22}(\theta) \end{bmatrix}$$

denote the partitions of $I(\theta)$ and $I(\theta)^{-1}$ corresponding to the partition of $\theta = (\theta_1, \underline{\theta}_2^{\mathsf{T}})^{\mathsf{T}}$ with $\underline{\theta}_2 = (\theta_2, \ldots, \theta_d)^{\mathsf{T}}$. Thus when θ_1 is the parameter of interest and $\underline{\theta}_2 = (\theta_2, \ldots, \theta_d)^{\mathsf{T}}$ are nuisance parameters, the *effective information* $I_{11}^*(\theta)$ for θ_1 is

$$I_{11}^*(\theta) = I_{11} - I_{12}I_{22}^{-1}I_{21} = E_\theta(\dot{\mathbb{l}}_1^{*2}), \tag{3.3}$$

where the *efficient score function* $\dot{\mathbb{l}}^*$ for θ_1 is

$$\dot{\mathbb{l}}_1^* \equiv \dot{\mathbb{l}}_1 - I_{12}I_{22}^{-1}\underline{\dot{\mathbb{l}}}_2 = \dot{\mathbb{l}}_1 - \Pi(\dot{\mathbb{l}}_1 \mid [\underline{\dot{\mathbb{l}}}_2]) \tag{3.4}$$

and the *efficient influence curve* $\tilde{\mathbb{l}}_1$ for estimation of θ_1 is

$$\tilde{\mathbb{l}}_1 = I_{11}^*(\theta)^{-1}\dot{\mathbb{l}}_1^*, \tag{3.5}$$

so that

$$E_\theta(\tilde{\mathbb{l}}_1^2) = I_{11}^*(\theta)^{-1} = I^{11}(\theta).$$

It is easily seen that the effective information I_{11}^* for θ_1 is just the squared length of the component $\dot{\mathbb{l}}_1^*$ of $\dot{\mathbb{l}}_1$ which is orthogonal to $\dot{\mathbb{l}}_2, \ldots, \dot{\mathbb{l}}_d$ in $L_2(P_\theta)$: in other words, the efficient score function is obtained by subtracting from $\dot{\mathbb{l}}_1$ its projection $\Pi(\dot{\mathbb{l}}_1 \mid [\underline{\dot{\mathbb{l}}}_2]) = I_{12}I_{22}^{-1}\underline{\dot{\mathbb{l}}}_2$ on the space $[\underline{\dot{\mathbb{l}}}_2]$ spanned by $\dot{\mathbb{l}}_2, \ldots, \dot{\mathbb{l}}_d$ in $L_2(P_\theta)$.

If the nuisance parameters $\underline{\theta}_2 = (\theta_2, \ldots, \theta_d)^{\mathsf{T}}$ are known, the bound (2) may be replaced by

$$n\mathrm{Var}_\theta[\hat{\theta}_1] \geqslant \frac{1}{I_{11}(\theta)}, \tag{3.6}$$

and, of course,

$$I_{11}(\theta) \geqslant I_{11}^*(\theta) = I_{11} - I_{12}I_{22}^{-1}I_{21}$$

where equality holds if and only if

$$I_{12} = I_{21}^{\mathsf{T}} = 0 \text{ or iff } \dot{\mathbf{l}}_1 \perp \dot{\mathbf{l}}_2, \ldots, \dot{\mathbf{l}}_d \text{ in } L_2(P_\theta). \tag{3.7}$$

Thus lack of knowledge of $\theta_2 \equiv (\theta_2, \ldots, \theta_d)^{\mathsf{T}}$ decreases the information for θ_1 unless (7) holds; in this case the lower bounds (2) and (6) agree, suggesting that θ_1 can be estimated as well when θ_2 is unknown as when θ_2 is known. This possibility was recognized by STEIN [73] in a paper which initiated the theory of *adaptive estimation.*

3.2. Hajek - Le Cam lower bounds

Two different but closely related asymptotic formulations of the classical Cramér - Rao lower bounds have proved useful: One is the convolution-type representation theorem of HAJEK [32] and LE CAM [48] which has been further developed and applied by BERAN [3], [4] and MILLAR [61]. The other is the local asymptotic minimax approach; see HAJEK [33] for a nice exposition and history, MILLAR [60], and LE CAM [49] for additional remarks.

Both types of lower bounds are formulated in terms of *locally asymptotically normal families*: Suppose that $\underline{X} = (X_1, \ldots, X_n) \cong P_{n,\theta}$ has density $p_n(\cdot, \theta)$, $\theta \in \Theta \subset \mathbb{R}^d$, and set

$$\mathbf{l}_n(\theta) = \log p_n(\underline{X}, \theta).$$

If $\theta_n \equiv \theta + hn^{-1/2}$, so that

$$\mathbf{l}_n(\theta_n) - \mathbf{l}_n(\theta) = \log[p_n(\underline{X}, \theta_n) / p_n(\underline{X}, \theta)],$$

then $\mathbf{P} \equiv \{P_{n,\theta} : \theta \in \Theta\}$ is *locally asymptotically normal* (LAN) at θ if there is a vector of $L_2(P_\theta)$ functions $\underline{\dot{\mathbf{l}}}_n(\theta)$ and a nonsingular matrix $I(\theta)$ such that, with

$$\mathbf{l}_n(\theta_n) - \mathbf{l}_n(\theta) = \underline{\dot{\mathbf{l}}}_n(\theta)^{\mathsf{T}} h - \tfrac{1}{2} h^{\mathsf{T}} I(\theta) h + R_n(\theta, h), \tag{3.8}$$

it follows that, in $P_{n,\theta}$-probability,

(i) $R_n(\theta, h) \to_p 0$ uniformly on bounded h-sets, and

(ii) $\underline{\dot{\mathbf{l}}}_n(\theta) \to_d N(0, I(\theta))$.

Thus $\mathbf{l}_n(\theta_n) - \mathbf{l}_n(\theta) \to_d N(-\frac{1}{2}\sigma^2, \sigma^2)$ with $\sigma^2 = h^{\mathsf{T}} I(\theta) h$. In 'regular families' $\mathbf{P}$ (with iid observations) $\underline{\dot{\mathbf{l}}}_n(\theta) = n^{-1/2} \Sigma_{i=1}^n \dot{\mathbf{l}}(\theta, X_i)$ where $\dot{\mathbf{l}}$ is the scores vector (for $n = 1$) and $I(\theta)$ is the information matrix.

Because our interest here in the parametric component θ of a semiparametric model $\mathbf{P} = \{P_{\theta,G}\}$, we formulate versions of the convolution and asymptotic minimax bounds for the first component θ_1 of θ.

A sequence of estimators T_{1n} of θ_1 is *regular* at θ if, under P_{θ_n}

$$\sqrt{n}(T_{1n} - \theta_{1n}) \to_d T_1$$

for every $\theta_n = \theta + n^{-1/2} h$ where the distribution $\mathbf{L}(T_1)$ of T_1 does not depend on h.

THEOREM 1 (HAJEK, 1970). *Suppose that* $\mathbf{P}$ *is LAN at* θ *and that* T_{1n} *is a regular estimator with limit distribution* $\mathbf{L}(T_1)$. *Then*

$$T_1 \cong Z_1 + W_1 \tag{3.9}$$

where $Z_1 \cong N(0, 1 / I_{11}^*(\theta))$, $I_{11}^*(\theta)$ *is as in* (3), *and* W_1 *is independent of* Z_1.

Thus any regular estimator T_{1n} of θ_1 must have a limit distribution which is at least as dispersed as $N(0,1/I_{11}^*(\theta))$, and it makes sense to call a regular estimator T_{1n} asymptotically efficient if $\sqrt{n}(T_{1n}-\theta_{1n})$ converges in distribution to Z_1; i.e. if $W_1=0$ in (9).
Now suppose that $w:\mathbb{R}^1\to\mathbb{R}^+$ satisfies:
(i) $w(x)=w(-x)$ for all $x\in\mathbb{R}^1$;
(ii) $w(0)=0$, $w(x)$ increases in $x\geq 0$;
(iii) $Ew(\sigma Z)<\infty$ for all $\sigma>0$ where $Z\cong N(0,1)$.

THEOREM 2 (HAJEK, 1972). *Suppose that* $\mathbf{P}$ *is LAN at* θ *and that w satisfies (i) - (iii). Then, for any estimator* T_{1n} *of* θ_1,

$$\lim_{M\to\infty}\liminf_{n\to\infty}\sup_{\sqrt{n}|\theta_n-\theta|\leq M} E_{\theta_n} w(\sqrt{n}(T_{1n}-\theta_{1n}))\geq Ew(Z_1) \tag{3.10}$$

where $Z_1\cong N(0,1/I_{11}^*(\theta))$ *as in theorem 1.*

If the uniformity in h in (i) of the definition of a LAN family is relaxed to just pointwise convergence, then theorems 1 and 2 continue to hold, but the bounds may not be attainable. Furthermore, if attention is restricted to regular estimates, then (10) holds without the supremum on the lefthand side.

3.3. Bounds for semiparametric models
The Hajek-Le Cam convolution and asymptotic minimax bounds stated above for a parametric model $\mathbf{P}_0$ continue to hold in a wide range of regular nonparametric and semiparametric models. All of the extensions make use, in some form, of the *tangent space* $\dot{\mathbf{P}}$ (at (θ,G)) for the model $\mathbf{P}$. For a parametric model $\mathbf{P}_0$ the tangent space $\dot{\mathbf{P}}_0$ (at $\theta\in\Theta$) is just the linear subspace $[\dot{\mathbf{l}}_1,\ldots,\dot{\mathbf{l}}_d]$ of $L_2(P_\theta)$ spanned by $\dot{\mathbf{l}}_1,\ldots,\dot{\mathbf{l}}_d$. For a semiparametric model $\mathbf{P}=\{P_{\theta,G}:\theta\in\Theta\subset\mathbb{R}^d,G\in\mathbf{G}\}$, the tangent space $\dot{\mathbf{P}}\subset L_2(P_{\theta,G})$ is simply the set of all possible score functions of one-dimensional regular parametric submodels (at (θ,G)).

For $\theta_0\in\Theta,G_0\in\mathbf{G}$, let $\mathbf{P}_\theta$ and $\mathbf{P}_G$ denote the submodels of $\mathbf{P}$ with $G=G_0$ and $\theta=\theta_0$ respectively:

$$\mathbf{P}_\theta\equiv\{P_{\theta,G_0}\in\mathbf{P}:\theta\in\Theta\},\quad \mathbf{P}_G\equiv\{P_{\theta_0,G}\in\mathbf{P}:G\in\mathbf{G}\}.$$

If $\dot{\mathbf{P}}_\theta$ and $\dot{\mathbf{P}}_G$ denote the corresponding tangent spaces, then $\dot{\mathbf{P}}_\theta\oplus\dot{\mathbf{P}}_G\subset\dot{\mathbf{P}}$ and typically equality holds. Here $\dot{\mathbf{P}}_G$ plays the role for estimation of θ, that $[\dot{\mathbf{l}}_2,\ldots,\dot{\mathbf{l}}_d]$ played for estimation of θ_1 in the parametric model $\mathbf{P}_0$, and the *efficient score function* for θ extending (4) is:

$$\dot{\mathbf{l}}_\theta^* = \dot{\mathbf{l}}_\theta - \Pi(\dot{\mathbf{l}}_\theta\,|\,\dot{\mathbf{P}}_G) \tag{3.11}$$

so that $\dot{\mathbf{l}}_\theta^*\perp\dot{\mathbf{P}}_G$ in $L_2(P_{\theta,G})$, and the *effective information* for θ in the model $\mathbf{P}$ is

$$I^*(\theta) = E_{\theta,G}(\dot{\mathbf{l}}_\theta^*\dot{\mathbf{l}}_\theta^{*\mathsf{T}}). \tag{3.12}$$

In the special case when $\dot{\mathbf{l}}_\theta^*=\dot{\mathbf{l}}_\theta\perp\dot{\mathbf{P}}_G$, then $I^*(\theta)=I(\theta)\equiv E_{\theta,G}(\dot{\mathbf{l}}_\theta\dot{\mathbf{l}}_\theta^{\mathsf{T}})$ and *adaptation to G* is possible; this is the situation emphasized by STEIN [73] and BICKEL [6].

Now versions of theorems 1 and 2 for the parametric component θ of the semiparametric model $\mathbf{P}$ continue to hold with θ_1 replaced by θ and $1/I_{11}^*(\theta)$ replaced by $I^*(\theta)^{-1}$ where $I^*(\theta)$ is given in (12); see KOSHEVNIK AND LEVIT [43], LEVIT [52], BEGUN et al. [1], and PFANZAGL [66], [67]. A complete treatment will be given in BICKEL, KLAASSEN, RITOV AND WELLNER [7].

4. Construction of asymptotically efficient estimates: two approaches

Suppose that $\mathbf{P}=\{P_{\theta,G}:(\theta,G)\in\Theta\times\mathbf{G}\}\equiv\{P_\theta:\theta=(\theta_1,\theta_2)\in\Theta_1\times\Theta_2\}$ with $\Theta_2=G$ is a 'regular' semiparametric model. A first stage in analyzing the model is to calculate scores for θ and information lower bounds as outlined in Section 3 above if possible. A second step is to construct estimators $(\bar{\theta}_n,\bar{G}_n)$ which are $\sqrt{n}$-consistent. A third stage is to find estimators $(\hat{\theta}_n,\hat{G}_n)$ of (θ,G) which are efficient in the sense that they actually *achieve* the information lower bounds (perhaps in the weakened sense of convergence in distribution for fixed (θ,G) rather than locally uniformly as required by the definition of regular estimates given in Section 3).

Two classical methods of constructing asymptotically efficient estimators $\hat{\theta}_n$ in regular parametric models are the methods of maximum likelihood estimation and Bayes estimation; see Lehmann [51] and Ibragimov and Has'minskii [38], though, as Lehmann makes clear, the emphasis in likelihood estimation, even in parametric models, should be on the scores and score equations rather than on maximizing likelihoods per se since the scores often lead to efficient estimates even when likelihoods themselves are unbounded.

Our aim here is to outline two useful approaches to the construction of asymptotically efficient estimates of the parametric part θ of a semiparametric model $\mathbf{P}$.

4.1. Method 1: Efficient score equation

Suppose that it is possible to calculate the *efficient score function* $\dot{\mathbf{l}}_1^*$ for θ_1,

$$\dot{\mathbf{l}}_1^* = \dot{\mathbf{l}}_1 - I_{12}I_{22}^{-1}\dot{\mathbf{l}}_2 = \dot{\mathbf{l}}_1 - \Pi(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{P}}_{\theta_2})$$

and the *effective information*

$$I_{11}^*(\theta) = E_\theta(\dot{\mathbf{l}}_1^{*2}).$$

Furthermore, suppose that $\bar{\theta}_n$ is a $\sqrt{n}$-consistent estimator of θ, $\sqrt{n}(\bar{\theta}_n-\theta)=O_P(1)$. Then define $\hat{\theta}_{1n}$ to be either a solution of the *efficient score equation*

$$\sum_{i=1}^{n}\dot{\mathbf{l}}_1^*(\hat{\theta}_{1n},\bar{\theta}_{2n},X_i)=0,$$

or a one-step approximation thereof:

$$\hat{\theta}_{1n}=\bar{\theta}_{1n}+\frac{\frac{1}{n}\sum_{i=1}^{n}\dot{\mathbf{l}}_1^*(\bar{\theta}_n,X_i)}{I_{11}^*(\bar{\theta}_n)} \tag{4.1}$$

$$=\bar{\theta}_{1n}+\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{l}}_1(\bar{\theta}_n,X_i)$$

where $\tilde{\mathbf{l}}_1$ is the efficient influence curve for θ_1, see (3.5). Additional smoothing may also be required in forming the sums in (1), but we have omitted it here for simplicity. Once an efficient estimator $\hat{\theta}_{1n}$ of θ_1 is found, method 2 can often be used to construct an efficient estimator of θ_2.

While no general theorem yet exists, the estimator $\hat{\theta}_{1n}$ defined above (or variations thereon involving suitable smoothing and truncation) has been shown to be asymptotically efficient in several important problems, a notable example being the errors in variables models studied by Bickel and Ritov [9]. Roughly speaking, the fact that $\dot{\mathbf{l}}_1^*$ is orthogonal to $\dot{\mathbf{l}}_2,\ldots,\dot{\mathbf{l}}_d$, the scores for θ_2, permits the use of an inefficient estimator $\bar{\theta}_{2n}$ to estimate out the 'nuisance parameter' θ_2. This should be contrasted with solving (or approximating by a one-step solution)

$$\sum_{i=1}^{n}\dot{\mathbf{l}}_1(\theta_1,\bar{\theta}_{2n})=0$$

for θ_1, a method which is known to produce inefficient estimates of θ_1 in general; see e.g. GONG AND SAMANIEGO [30].

The main drawback of the method is that it requires calculation of the efficient score function $\dot{\ell}_1^*$. Thus the method depends heavily on being able to calculate projections onto $[\dot{\ell}_2] = \dot{\mathbf{P}}_{\theta_2} = \dot{\mathbf{P}}_G$, which often necessitates calculation of the inverse of the information operator $\dot{\ell}_2^T \dot{\ell}_2 = I_{22}$. When $\dot{\ell}_1^* = \dot{\ell}_1$ so $\dot{\ell}_1$ is orthogonal to $[\dot{\ell}_2] = \dot{\mathbf{P}}_{\theta_2}$, then 'adaptation' with respect to $\theta_2 = G$ is possible, and method 1 becomes essentially the method used to construct efficient estimates in this case; see e.g. STONE [75] and BICKEL [6].

4.2. Method 2: Efficient estimation of θ_2 for known θ_1

Now suppose that an *efficient* estimate $\tilde{\theta}_{2n}$ of θ_2 is available if θ_1 is *known*. We denote this estimator by $\tilde{\theta}_{2n}(\theta_1)$ because it depends on the 'known' value of θ_1. Substitution of this estimate of θ_2 into the ordinary score for θ_1 (as if θ_2 were known and equal to $\tilde{\theta}_{2n}$) yields the 'condensed' or 'concentrated' score equation

$$\sum_{i=1}^{n} \dot{\ell}_1(\theta_1, \tilde{\theta}_{2n}(\theta_1), X_i) = 0$$

which we can solve for $\theta_1 \equiv \hat{\theta}_1$. Or, if $\bar{\theta}_{1n}$ is a $\sqrt{n}$-consistent estimate of θ_1, a one-step approximation thereof:

$$\hat{\theta}_{1n} = \bar{\theta}_{1n} + \frac{\frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_1(\bar{\theta}_{1n}, \tilde{\theta}_{2n}(\bar{\theta}_{1n}), X_i)}{\frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_1^2(\bar{\theta}_{1n}, \tilde{\theta}_{2n}(\bar{\theta}_{1n}))}. \tag{4.2}$$

As in the case of (1), more smoothing may be needed in forming the sums in (2); we have omitted it here for simplicity. This is a frequently used device in parametric models, but the method is equally useful for semiparametric models. While no general results concerning the estimator (2) seem to be known, this method has been used by RITOV [71] to construct efficient estimates for censored regression models.

5. PROBLEMS

Statisticians have a large, well-stocked tool-box for dealing with classical parametric models, and a growing companion set of tools for handling completely nonparametric models. The choice of tools for dealing with the rich middle ground of semiparametric models is, however, still relatively limited, and the few available tools are not all well suited for the job. Many important problems remain. Here is a partial list:

(a). *Calculation of lower bounds.* If the projection $\Pi(\dot{\ell}_\theta \mid \dot{\mathbf{P}}_G)$ in Section 3 can be calculated, then so can the efficient score function $\dot{\ell}_\theta^*$, the effective information $I_{11}^*(\theta)$, and the efficient influence curve $\tilde{\ell}_1$. In many models this projection is simply a conditional expectation, and hence can be calculated easily; but in other models such as the dependent proportional hazards model of 2.E(b) the projection calculation is apparently intractable. More systematic methods, possibly involving iterative, numerical techniques, are needed.

(b). *Construction of efficient estimates.* HUANG [36] has made a preliminary study of method 1 outlined in Section 4, but general results concerning the asymptotic efficiency of methods 1 and 2, or variations thereof involving more smoothing, are still needed. Other methods including minimum Hellinger distance estimates, minimum Kullback-Leibler discrepancy estimators, and maximum-likelihood estimators obtained via EM-algorithms all need further development and sharpening in the context of semiparametric models. Efficient estimates are still unknown for many of the

models given in Section 2.

(c). *Identifiability and regularity criteria.* For many semiparametric models, further work on identifiability and conditions for regularity of submodels is still needed before work on estimation can get underway. For examples of such studies, see the papers by HECKMAN AND SINGER [35] and ELBERS AND RIDDER [25] concerning identifiability issues for the models of 2.E(b) and 2.E(c). Classical regularity investigations of translation and parametric models, which carry over to many group models are given by HAJEK [31], [33].

(d). *Hypothesis testing.* As yet no adequate theory of hypothesis testing exists for semiparametric models. One type of testing problem concerns testing hypotheses within a nested family of semiparametric models: for example, consider testing $\Lambda_2 = \gamma\Lambda_1$ for some $0<\gamma<\infty$ in the Clayton-Oakes model of example E(b). Or, of interest in survival analysis, test the assumption of a proportional hazards regression model against some general family of alternatives. Another rather different testing problem would involve testing non-nested semiparametric models against one another, e.g. a Cox-type regression model against a more classical linear regression model or perhaps a semiparametric mixed regression model.

(e). *Asymptotics for estimates based on smoothing.* Construction of efficient estimates for many of the models discussed above require smoothing techniques involving density or conditional expectation estimators. While the asymptotics for such smoothing processes are available, they need further development, study, and refinement to ease their systematic application to the construction of efficient estimates in a wide range of semiparametric models.

(f). *Robustness; connections and problems.* Efficient estimation in semiparametric models has many interesting connections with questions of robustness. Just as classical robustness theory has focused on neighborhoods of parametric models (often a one - sample location model), a generalization suggested by BICKEL and LEHMANN [8] concerns neighborhoods of semiparametric models, which they called 'nonparametric models with natural parameters'. For example, are the partial likelihood estimtors for the Cox proportional hazards model robust in some appropriate sense (with respect to the assumption of proportional hazards)? As more experience is gained with efficient estimates for semiparametric models, this more general type of robustness outlined by BICKEL and LEHMANN [8] can begin to be considered. Many challenging problems remain.

Acknowledgments: I have profited from several helpful discussions concerning semiparametric models with Peter Bickel. In particular, I learned of 'method 2' in Section 4 from him. I also owe thanks to Richard Gill for helpful comments concerning Sections 1 and 3. R.D. Martin suggested example 2D(f).

REFERENCES

1. J.M. BEGUN, W.J. HALL, W.M. HUANG and J.A. WELLNER (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist. 11,* 432 - 452.
2. S. BENNETT (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine 2,* 273-277.
3. R. BERAN (1977). Estimating a distribution function. *Ann. Statist. 5,* 400-404.
4. R. BERAN (1977). Robust location estimates. *Ann. Statist. 5,* 431-444.
5. P.K. BHATTACHARYA, H. CHERNOFF AND S.S. YANG (1983). Nonparametric estimation of the slope of a truncated regression. *Ann. Statist. 11,* 505-514.
6. P.J. BICKEL (1982). On adaptive estimation. *Ann. Statist. 10,* 647-671.
7. P.J. BICKEL, C.A.J. KLAASSEN, Y. RITOV AND J.A. WELLNER (1986). *Efficient and Adaptive Inference in Semiparametric Models,* forthcoming monograph, Johns Hopkins University Press, Baltimore.
8. P.J. BICKEL AND E.L. LEHMANN (1975). Descriptive statistics for nonparametric models. I.

Introduction. *Ann. Statist. 3,* 1038-1044.
9. P.J. BICKEL AND Y. RITOV (1984). *Efficient Estimation in the Errors in Variables Model,* preprint, Dept. of Statistics, University of California, Berkeley.
10. D. BRANSTON (1976). Models of single lane time headway distributions. *Transportation Science 10,* 125-148.
11. L. BREIMAN AND J. FRIEDMAN (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc. 80,* 580-619 (with discussion).
12. N.E. BRESLOW AND N.E. DAY (1980). *The Analysis of Case-Control Studies,* International Agency for Research on Cancer, Lyon.
13. J. BUCKLEY AND I. JAMES (1979). Linear regression with censored data. *Biometrika 66,* 429-436.
14. R.J. CARROLL (1984). *Adaptation for the Slope in Simple Logistic Regression with an Intercept in the Structural Errors in Variables Model,* preprint.
15. R.J. CARROLL (1984). *A General Technique for Computing Information Bounds in Errors in Variables Structural Models,* preprint.
16. D. CLAYTON (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika 65,* 141-151.
17. D. CLAYTON AND J. CUZICK (1985). Multivariate generalizations of the proportional hazards model. *J. Roy. Statist. Soc. Ser. A. 148,* 82-117 (with discussion).
18. D. CLAYTON AND J. CUZICK (1985). *An Approach to Inference for Rank-Regression Models with Right-Censored Data,* preprint.
19. D. CLAYTON AND J. CUZICK (1985). The semi-parametric Pareto model for regression analysis of survival times. *Bull. Int. Stat. Inst. 51, part 4,* 23.3.1 - 23.3.18; pp. 19-30 in this report.
20. S.R. COSSLETT (1981). Maximum likelihood estimation for choice-based samples. *Econometrica 49,* 1289-1316.
21. D.R. COX (1969). Some sampling problems in technology. N.L. JOHNSON AND H. SMITH, JR. (eds.). *New Developments in Survey Sampling,* 506-527, Wiley-Interscience, New York.
22. D.R. COX (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B 34,* 187-220.
23. D.R. COX AND E.J. SNELL (1981). *Applied Statistics: Principles and Examples,* Chapman and Hall, London.
24. K. DOKSUM (1985). *Partial Likelihood Methods in Transformation Models,* preprint, University of California, Berkeley.
25. C. ELBERS AND G. RIDDER (1983). True and spurious duration dependence: the identifiability of the proportional hazards model. *Review of Economic Studies 49,* 403-410.
26. R.F. ENGLE, C.W.J. GRANGER, J. RICE AND A. WEISS (1983). *Nonparametric Estimates of the Relation between Weather and Electricity Demand,* preprint, Department of Economics, University of California, San Diego.
27. J.H. FRIEDMAN AND W. STUETZLE (1981). Projection pursuit regression. *J. Amer. Statist. Assoc. 76,* 817-823.
28. R.D. GILL (1983). *Models for the Censored Data Matched Pairs Problem,* Unpublished manuscript, Centrum voor Wiskunde en Informatica, Amsterdam.
29. R.D. GILL AND J.A. WELLNER (1985). *Limit Theorems for Empirical Distributions in Selection Bias Models,* preprint, Dept. of Statistics, University of Washington.
30. G. GONG AND F.J. SAMANIEGO (1981). Pseudo maximum likelihood estimation: theory and applications. *Ann. Statist. 9,* 861-869.
31. J. HAJEK (1962). Asymptotically most powerful rank order tests. *Ann. Math. Statist. 33,* 1124-1147.
32. J. HAJEK (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. verw. Gebiete 14,* 323-330.

33. J. HAJEK (1972). Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berk. Symp. Math. Statist. Prob. 1,* 175-194, University of California Press, Berkeley, California.
34. R.Z. HAS'MINSKII AND I.A. IBRAGIMOV (1983). On asymptotic efficiency in the presence of an infinite dimensional nuisance parameter. K. ITO AND J.V. PROKHOROV (eds.). *Probability Theory and Mathematical Statistics, Fourth USSR - Japan Symposium, Lecture Notes in Mathematics, 1021,* 95-229, Springer - Verlag, Berlin.
35. J. HECKMAN AND B. SINGER (1984). A method for minimizing the impact of distributional assumptions in economic studies for duration data. *Econometrica 52,* 271-320.
36. W. HUANG (1984). *On Effective Score Estimation in Semiparametric Models,* preprint.
37. P.J. HUBER (1985). Projection pursuit. *Ann. Statist. 13,* 435-525 (with discussion).
38. I.A. IBRAGIMOV AND R.Z. HAS'MINSKII (1981). *Statistical Estimation: Asymptotic Theory,* Springer-Verlag, New York.
39. N.P. JEWELL (1982). Mixtures of exponential distributions. *Ann. Statist. 10,* 479-484.
40. N.P. JEWELL (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika 72,* 11-21.
41. N.P. JEWELL (1985). *Least Squares Estimation of the Slope of a Truncated Regression,* preprint, Department of Biostatistics, University of California, Berkeley.
42. J. KIEFER AND J. WOLFOWITZ (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Ann. Math. Statist. 27,* 887-906.
43. YU.A. KOSHEVNIK AND B.YA. LEVIT (1976). On a nonparametric analogue of the information matrix. *Theor. Prob. Appl. 21,* 738-753.
44. J. LAFLER AND T.D. KINMAN (1965). An RR Lyrae star survey with the Lick 20 - inch astrograph II. The calculation of RR Lyrae periods by the electronic computer. *Astrophysical J., Suppl. 11,* 216-222.
45. N. LAIRD (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc. 73,* 805-811.
46. D. LAMBERT AND L. TIERNEY (1984). Asymptotic efficiency of estimators of functionals of mixed distributions. *Ann. Statist. 12,* 1380-1387.
47. D. LAMBERT AND L. TIERNEY (1984). Asymptotic properties of maximum likelihood estimates in the mixed Poisson model. *Ann. Statist. 12,* 1388-1399.
48. L. LE CAM (1972). Limits of experiments. *Proc. Sixth Berkeley Symp. Math. Statist. and Prob. 1,* 245-261, University of California Press, Berkeley, California.
49. L. LE CAM (1984). Review of Ibragimov and Has'minskii (1981) and Pfanzagl (1982), *Bull. (New Series) Amer. Math. Soc. 11,* 391-400.
50. E.L. LEHMANN (1953). The power of rank tests. *Ann. Math. Statist. 24,* 23-43.
51. E.L. LEHMANN (1983). *Theory of Point Estimation,* Wiley, New York.
52. B.YA. LEVIT (1978). Infinite-dimensional informational lower bounds. *Theor. Prob. Applic. 20,* 723-740.
53. B. LINDSAY (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. *Philos. Trans. Roy. Soc. London Ser. A 296,* 39-665.
54. B. LINDSAY (1983). The geometry of mixture likelihoods, part I. *Ann. Statist. 11,* 86-94.
55. B. LINDSAY (1983). The geometry of mixture likelihoods, part II. *Ann. Statist. 11,* 783-792.
56. C.F. MANSKI AND S.R. LERMAN (1977). The estimation of choice probabilities from choice based samples, *Econometrica 45,* 1977-1988.
57. C.F. MANSKI AND D. MCFADDEN (1981). *Structured analysis of discrete data,* MIT Press.
58. J. MCDONALD (1983). Periodic smoothing of time series. *Project Orion Technical Report 017,* Department of Statistics, Stanford University, Stanford, California.
59. P.W. MILLAR (1979). Asymptotic minimax theorems for the sample distribution. *Z. Wahrsch. verw. Gebiete 48,* 233-252.

60. P.W. MILLAR (1983). The minimax principle in asymptotic statistical theory, *Proc. Ecole d'Ete St. Flour, Lecture Notes in Math. 976,* 75-265, Springer - Verlag, Berlin.
61. P.W. MILLAR (1985). Nonparametric applications of an infinite dimensional convolution theorem. *Z. Wahrsch. verw. Gebiete 68,* 545-556.
62. R. MILLER AND J. HALPERN (1982). Regression with censored data. *Biometrika 69,* 521-531.
63. R.G. MILLER (1976). Least squares regression with censored data. *Biometrika 63,* 449-464.
64. D. OAKES (1982). A model for association in bivariate survival data. *J. Roy. Statist. Soc. 44, Ser. B,* 412-422.
65. D. OAKES (1985). *Semiparametric Estimation in a Model for Association in Bivariate Survival Data,* preprint, Dept. of Statistics, University of Rochester (to appear in *Biometrika*)
66. J.PFANZAGL (1982). *Contributions to a General Asymptotic Statistical Theory, Lecture Notes in Statistics 13,* Springer - Verlag, New York.
67. J. PFANZAGL (1984). *A Remark on Semiparametric Models,* preprint, University of Cologne.
68. R.L. PRENTICE AND R. PYKE (1979). Logistic disease incidence models and case-control studies. *Biometrika 66,* 403-411.
69. R. PRENTICE AND S. SELF (1983). Asymptotic distribution theory for Cox-type regression models with general risk form. *Ann. Statist. 11,* 804-813.
70. G. RIDDER AND W. VERBAKEL (1983). *On the Estimation of the Proportional Hazards model in the Presence of Unobserved Heterogeneity,* preprint. (to appear in *J. Amer. Statist. Assoc.*)
71. Y. RITOV (1984). *Efficient and Unbiased Estimation in Nonparametric Linear Regression with Censored Data,* preprint, Department of Statistics, University of California, Berkeley.
72. A. SCHICK. (1984). *On adaptive estimation,* preprint.
73. C. STEIN (1965). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Prob. 1,* 187-195, University of California Press, Berkeley, California.
74. R.F. STELLINGWERF (1978). Period determination using phase dispersion minimization. *Astrophysical J. 224,* 953-960.
75. C.J. STONE (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist. 3,* 267-284.
76. C.J. STONE (1985). Additive regression and other nonparametric models. *Ann. Statist. 33,* 689-705.
77. R. TIBSHIRANI (1982). Censored data regression with projection pursuit. *Project Orion Technical Report 013,* Department of Statistics, Stanford University.
78. R. TIBSHIRANI (1983). Non-parametric estimation of relative risk. *Project Orion Technical Report 022,* Department of Statistics, Stanford University.
79. Y. VARDI (1983). Nonparametric estimation in the presence of length bias. *Ann. Statist. 10,* 616-620.
80. Y. VARDI (1985). Empirical distributions in selection bias models. *Ann. Statist. 13,* 178-203.
81. G. WAHBA (1984). Partial spline models for the semiparametric estimation of functions of several variables. *Seminar Proceedings, Japan - USSR Joint Seminar on the Statistical Analysis of Time Series.*

The Semi-Parametric Pareto Model for Regression Analysis of Survival Times

David Clayton
Department of Community Health, Clinical Science Building
University of Leicester
Leicester Royal Infirmary
P.O. Box 65
Leicester LE2 7LC
England

Jack Cuzick
Imperial Cancer Research Fund Laboratories
Department of Mathematics, Statistics and Epidemology
P.O. Box 123
Lincoln's Inn Fields
London WC2A 3PX
England

The Pareto distribution may be derived as a gamma mixture of exponential distributions and arises in modelling survival times in heterogeneous populations. Cox's proportional hazard model is a semi-parametric generalization of the exponential regression model, and we discuss the equivalent generalization of the Pareto model. This model predicts convergent hazards and leads to the general family of efficient score tests described by HARRINGTON AND FLEMING [16]. A special case is the proportional odds model discussed by BENNETT [4]. We review approaches to point and interval estimation in such models and discuss extensions of the model to describe associated failure times.

1. THE MODEL.

The exponential regression model for survival times (PRENTICE [28]) holds that an individual characterised by covariate vector z experiences constant hazard which depends only on the value of z. When this dependence is log-linear, this takes the form

$$\lambda(t \mid z) = \lambda_0 \exp(\beta^T z) \tag{1.1}$$

Alternatively, the model may be written as a classical linear model for log failure times.

$$\log(t) = \mu - \beta^T z + \epsilon \tag{1.2}$$

where $\mu = -\log\lambda_0$ and ϵ is a standard extreme value (s.e.v.) variate.

The proportional hazards model introduced by Cox [11] generalizes (1.1) to allow the baseline hazard to vary with time, i.e.

$$\lambda(t \mid z) = \lambda_0(t) \exp(\beta^T z) \tag{1.3}$$

where $\lambda_0(t)$ is some unknown baseline hazard function. The equivalent linear model for failure times may now be written

$$g(t) = -\beta^T z + \epsilon \tag{1.4}$$

where g is a non-decreasing function and ϵ is, once again, an s.e.v. variate. The equivalence of (1.3) and (1.4) is easily established by noting that

$$g(t) = \log \int_0^t \lambda_0(u)du = \log\Lambda_0(t) \tag{1.5}$$

Extension of the simple exponential model (1.1) to include unexplained population heterogeneity involves introducing an unknown covariate ω

$$\lambda(t \mid z,\omega) = \lambda_0 \exp(\beta^T z + \omega) = \lambda_0 \, \xi \exp(\beta^T z) \tag{1.6}$$

where ξ is a random variate with distribution function $\Phi(\xi)$, which is often referred to as the 'frailty' (VAUPEL, MANTON AND STALLARD [31]). In the simplest case, we assume the frailty distribution to be a gamma distribution with unit mean and variance γ. With this assumption, it is easily shown that the unconditional survivor function is

$$F(t \mid z) = \int_\xi F(t \mid z,\xi) \, d\Phi(\xi) = (1 + \gamma\theta t)^{-1/\gamma} \tag{1.7}$$

where θ is the relative risk term $\exp(\beta^T z)$. This is the Pareto distribution. It has hazard function $\theta(1+\gamma\theta t)^{-1}$ which is monotone decreasing.

The alternative representation of (1.6) as a linear model for log survival times is as (1.2) but now the random error term, ϵ, is the log of a beta variate of the second kind with parameters 1 and γ^{-1}. An important special case is when $\gamma=1$ when the error distribution is logistic. Equation (1.2) then gives the log-logistic model discussed by BENNETT [3].

The generalization of the Pareto model to allow time-dependent baseline hazard follows naturally. In particular, (1.6) becomes

$$\lambda(t \mid z,\omega) = \lambda_0(t) \, \xi \exp(\beta^T z) \tag{1.8}$$

and (1.7) becomes

$$F(t \mid z) = \{1 + \gamma\theta\Lambda_0(t)\}^{-1/\gamma} \tag{1.9}$$

It may be shown that the ratio of the hazard functions between subgroups with relative risk functions θ_1 and θ_2 is initially θ_1 / θ_2 but converges to 1 as $t \to \infty$. The linear model representation is as (1.4) and (1.5), but with a different error distribution.

In this more general setup, the special case of $\gamma=1$ becomes the proportional odds model, see BENNETT [4]. This bears the same relationship to the log-logistic regression model as does the proportional hazards model to the exponential regression model. The 'proportional odds' property arises from (1.9), which becomes, with $\gamma=1$,

$$\{1 - F(t \mid z)\} \; / \; F(t \mid z) = \theta \, \Lambda_0(t) \tag{1.10}$$

so that the relative odds of survival for two individuals with relative risk functions θ_1 and θ_2 remains constant over time (at θ_1 / θ_2).

CLAYTON AND CUZICK [9] discuss a further generalization of (1.8) in which the random effects ('frailties') are shared between related individuals. That paper concentrated primarily upon estimation of the parameter, γ, of the frailty distribution. Here we discuss estimation of the regression coefficients, β, in the simpler case of unshared frailties with γ a known constant.

2. THE SIMPLE PARAMETRIC MODEL

In this section we discuss maximum likelihood estimation of the regression coefficients, β, in the case where $\Lambda_0(t)$ is a known function. This problem arises in epidemiological cohort studies of mortality when $\Lambda_0(t)$ may be calculated from population vital statistics. For simplicity, in this and later sections we shall assume the variance, γ, of the frailty distribution also to be known. Simultaneous estimation of γ and β would, however, present no difficulties in principle.

We observe times t_i, $i=1,...,N$, in individuals characterized by covariate vectors $\{z_i\}$. The observed times are either known failure times or times of (independent) right censoring. We shall use the indicator variables $\{d_i\}$ to denote failure or censoring (with the values 1 and 0 respectively). We shall

denote the transformed times $\Lambda_0(t_i)$ by y_i. As above, we denote the relative risk function, $\exp(\beta^T z_i)$, by θ_i.

If we write F(y) for the standard Pareto survivor function (1.7) with relative risk $\theta = 1$, and f(y) for the corresponding density, then the log-likelihood is given by

$$\ell = \sum_i \left[d_i \log [\theta_i f(\theta_i y_i)] + (1-d_i)\log F(\theta_i y_i) + d_i \log \lambda_0(t_i) \right]. \tag{2.1}$$

The last term arises from the Jacobean $\frac{\partial y}{\partial t}$. The first and second derivatives are, respectively,

$$\frac{\partial \ell}{\partial \beta} = \sum_i \left[d_i - \theta_i y_i \zeta(\theta_i y_i) \right] z_i \tag{2.2}$$

and

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta^T} = - \sum_i \left[\theta_i y_i \zeta(\theta_i y_i) + (\theta_i y_i)^2 \zeta'(\theta_i y_i) \right] z_i z_i^T \tag{2.3}$$

where the function $\zeta(y) \equiv \zeta(y,d)$ is defined as

$$\zeta(y) = - f'(y)/f(y) = \phi(y), \text{ for } \textit{uncensored observations } (d=1), \tag{2.4}$$

and

$$\zeta(y) = - F'(y)/F(y) = \lambda(y), \text{ for } \textit{censored observations } (d=0). \tag{2.5}$$

These results hold in general for other error distributions, and were given by PRENTICE [29]. In particular, (2.2) and (2.3) may be used in the construction of efficient score tests for $H_0: \beta = \beta_0$. For the Pareto model considered here,

$$\zeta(y) = (1 + \gamma d) \; / \; (1 + \gamma y). \tag{2.6}$$

We may also note that the term $\theta_i y_i \zeta(\theta_i y_i)$ in (2.2) may be rewritten

$$\theta_i y_i \zeta(\theta_i y_i) = (d_i + \gamma^{-1}) \pi_i \tag{2.7}$$

where

$$\pi_i / (1 - \pi_i) = \gamma \theta_i y_i \tag{2.8}$$

so that for ML estimation of β, the problem is numerically identical to that presented by logistic regression analysis for a binary response variable, d, with binomial denominator $(d + \gamma^{-1})$. The likelihood may be maximized using the computer program GLIM (BAKER AND NELDER [1]). This has been noted by CLAYTON AND CUZICK [9] and by BENNETT [3].

3. PARAMETRIC MODELS FOR $\Lambda_0(t)$.

We now consider a more general model in which $\Lambda_0(t)$ is some unknown function, $\Lambda_0(t;\alpha)$ depending on a (possibly vector) parameter α. Estimation of α is equivalent to finding the transformation in the parametric family $y = \Lambda_0(t;\alpha)$ such that the transformed times, $\{y_i\}$, are best represented by the model described above. There is a strong connection between the generalization described in this section and the generalization of the classical linear model proposed by BOX AND COX [5].

For known β, ML estimation of α involves solving the equations obtained by setting the first derivatives

$$\frac{\partial \ell}{\partial \alpha} = \sum_i \left[d_i \frac{\partial}{\partial \alpha} \log \lambda_0(t_i;\alpha) + \theta_i \zeta(\theta_i y_i) \frac{\partial}{\partial \alpha} \log \Lambda_0(t_i;\alpha) \right] \tag{3.1}$$

to zero (notation is as in section 2).

It is interesting to consider joint ML estimation of α and β carried out in a two step recursion as

follows:

- Step 1: Maximize the likelihood with respect to α, treating the current estimates of β as known constants, i.e. find the best data transformation for the current regression model, and

- Step 2: Maximize the likelihood with respect to β, treating the current estimates of α as known, i.e. fit the regression model for the current transformation of the data.

Although this may not be the most efficient numerical method for ML estimation, it is useful in establishing parallels with the semi-parametric approach discussed later. If we denote the second derivatives of the log-likelihood with respect to α by $L_{\alpha\alpha}$, those with respect to β (given by equation (2.3)) by $L_{\beta\beta}$, and those with respect to both α and β by $L_{\alpha\beta}$, then a consistent estimator of the variance-covariance matrix of the ML estimates of β is given by the appropriate portion of the inverse of the observed information matrix, i.e. by

$$\mathrm{Var}(\hat{\beta}) = [\ -L_{\beta\beta} + L_{\alpha\beta}^{\mathsf{T}} L_{\alpha\alpha}^{-1} L_{\alpha\beta}\ ]^{-1} \tag{3.2}$$

where $L_{\alpha\alpha}$ etc. are evaluated at the ML estimates of α and β. Note that the second term in (3.2) causes an increase in variance over that which would apply if α were known as in section 2. This term represents the loss in information incurred by estimation of $\Lambda_0(t)$. If we write ℓ_i for the contribution of the i-th individual to the log-likelihood (2.1), and B for the matrix of second derivatives

$$B_{ij} = \partial^2 \ell_i / \partial y_i \partial \beta_j = -[\ \theta_i \zeta(\theta_i y_i) + \theta_i^2 y_i \zeta'(\theta_i y_i)\ ]\ z_{ij} \tag{3.3}$$

then the correction term in (3.2) may be rearranged and we have

$$\mathrm{Var}(\hat{\beta}) \approx [\ -L_{\beta\beta} - B^{\mathsf{T}} \Sigma B\ ]^{-1} \tag{3.4}$$

where

$$\Sigma_{ij} = (\partial y_i / \partial \alpha^{\mathsf{T}})\ (-L_{\alpha\alpha})^{-1}\ (\partial y_j / \partial \alpha). \tag{3.5}$$

Now, $(-L_{\alpha\alpha})^{-1}$ estimates the sampling variance of the ML estimates of α for known β, so that Σ may be thought of as an estimate of the variance-covariance matrix of the transformed values $\{y_i\}$, for known $\{t_i\}$ and β, arising out of errors of estimation of α. We shall encounter an analogous expression to (3.4) in the next section when we discuss the semi-parametric approach.

4. The semi-parametric approach.

We now consider a semi-parametric approach in which we attempt to estimate β without explicit parametric assumptions concerning the transformation $\Lambda_0(t)$. We assume only that $\Lambda_0(t)$ is a monotone increasing function.

In the case of the proportional hazards model, this is achieved by recourse to 'partial likelihood' arguments (see Cox [11], [12]), but this approach does not seem to be more generally useful. Instead we return to the proposal of Kalbfleisch and Prentice [18] and [19], who suggested that invariance arguments lead to consideration of the marginal likelihood based upon the generalized rank vector of the $\{t_i\}$. If we assume the sample is ordered so that $t_1 < t_2 \ldots < t_N$ then, for uncensored data, this marginal likelihood is given by

$$L_m = \int_{y_1 < y_2 \ldots < y_N} \cdots \int \prod_i \{\theta_i f(\theta_i y_i)\ \mathrm{d}y_i\} \tag{4.1}$$

where θ_i and $f(y)$ are as in in section 2. Kalbfleisch and Prentice suggested that this marginal likelihood may be extended to allow for right-censored observations by extending the range of integration to cover the entire domain of the underlying vector y consistent with the incomplete rank information. Thus expressions such as (4.1) are summed over all possible underlying rank vectors. Then, if $j_1, j_2 \ldots, j_M$ represent the indices of the uncensored times ($j_1 < j_2 \ldots < j_M, M \leq N$) and J_i represents the set of indices of observations censored between the i-th and (i+1)-th uncensored failure times,

Kalbfleisch and Prentice showed that (4.1) generalizes to

$$L_m = \int_{y_{j1}<y_{j2}\cdots<y_{jM}} \cdots \int \prod_i \left[\theta_{j_i} f(\theta_{j_i} y_{j_i}) \prod_{k \in J_i} F(\theta_k y_{j_i}) \, \mathrm{d}y_{j_i} \right]. \tag{4.2}$$

Note that the integrand of (4.2) is the parametric likelihood whose logarithm is given by (2.1).

For the proportional hazards model, (4.2) corresponds to the partial likelihood. It is only a true likelihood, however, in the highly artificial case of 'progressive type II censoring' in which the positions of the censored observations in the rank vector are determined in advance of carrying out the experiment. Nevertheless, one would expect that, like partial likelihood, (4.2) will behave asymptotically as a true likelihood under more realistic assumptions about the censoring mechanism.

PRENTICE [29] used this likelihood to derive a series of efficient score tests similar to those proposed by PETO AND PETO [23]. More recently, the asymptotic properties of these tests have been demonstrated by CUZICK [14] using a general random censoring model.

For the purposes of estimation, however, this marginal likelihood is quite intractable except in the special case of proportional hazards. CLAYTON AND CUZICK [9] proposed an approximate method for maximum marginal likelihood estimation using a quasi-EM algorithm. If we denote the generalized rank vector for the censored survival times by $R = (R_1, \ldots, R_N)$, where

$$R_i = (j_i, J_i) \tag{4.3}$$

then the algorithm is a two-step recursion:

- Step 1: (E-step) Evaluate the scores $\{\bar{y}_i\}$ using the current estimates of the regression coefficients, β (and γ), where

$$\bar{y}_i = E(y_i \mid R; \beta, \gamma) \tag{4.4}$$

- Step 2: (M-step) Maximize with respect to β the function given by substituting the scores $\{\bar{y}_i\}$ into the parametric log-likelihood (2.1).

We believe it will be possible to prove that this procedure leads to consistent estimates which asymptotically approach the maximum marginal likelihood (MML) estimator. The true EM algorithm, see DEMPSTER, LAIRD AND RUBIN [15], would involve, at the M-step, solution of the normal equations obtained by setting the right-hand side of (2.2) to zero after first, at the E-step, having replaced the terms $\{\theta_i y_i \zeta(\theta_i y_i)\}$ by their expectations given the rank vector and the current estimates of β and γ. In the case of the proportional hazards model ($\gamma = 0$), $\zeta(\theta_i y) = 1$ for every i and for all y and the method is a true EM algorithm. In this case, also, the scores $\{\bar{y}_i\}$ may be explicitly calculated, and we have a novel computational method for the Cox model (CLAYTON AND CUZICK [10]). In other cases, the expectations cannot be calculated exactly, and we suggest approximations in the next section.

An estimate of the variance of the MML estimator of β is given by (minus) the inverse of the second derivative matrix of the log of the marginal likelihood. If $L_{\beta\beta}$ is the second derivative matrix of the parametric log-likelihood (2.3), evaluated at the final scores $\{\bar{y}_i\}$ and at the MML estimates of β, then, by differentiation of the Taylor expansion of the marginal log-likelihood about $\{\bar{y}_i\}$, we may obtain the following asymptotic variance estimator

$$\mathrm{Var}(\hat{\beta}) \approx [\, - L_{\beta\beta} - B^{\mathsf{T}} \Sigma B \,]^{-1} \tag{4.5}$$

where $\Sigma_{ij} = \mathrm{Cov}\{y_i, y_j \mid R; \beta, \gamma\}$ and B is as in (3.4).

A similar expression may be shown to hold for the true EM-algorithm. Using results from DEMPSTER, LAIRD AND RUBIN [15], the inverse of the information matrix may be shown to be as in (4.5) but the correction term $B^{\mathsf{T}} \Sigma B$ is replaced by the asymptotically equivalent expression $Z^{\mathsf{T}} \Gamma Z$ where Z is the matrix with rows $\{z_i\}$ and Γ is the conditional covariance matrix of the $\{\theta_i y_i \zeta(\theta_i y_i)\}$.

5. CALCULATION OF SCORES.

The scores $\{\bar{y}_i\}$ of (4.4) cannot in general be explicitly calculated. Nevertheless, adequate approximations can be obtained. Here we indicate a possible approach and give the main results and a heuristic argument. More work is needed to justify these calculations rigorously.

Initially we ignore censoring and assume that we observe random variates $\{y_i\}$, i=1,...,N independently distributed with densities $\{f_i(y)\}$. Without loss of generality, assume that the y_i are ordered so that the rank vector, R, carries the information that $y_1 < y_2 \cdots < y_N$. We require $E(y_i \mid R)$ or an adequate approximation. The densities are assumed to possess sufficiently many derivatives to justify the expansions below.

We shall write

$$m_i = (y_{i-1} + y_{i+1})/2 \tag{5.1}$$
$$r_i = y_{i+1} - y_{i-1}$$

and, expanding about $y = m_i$, we obtain

$$E(y_i \mid R, y_{i-1}, y_{i+1}) = m_i - \frac{1}{12} r_i^2 \phi_i(m_i) + \mathcal{O}(r_i^4) \tag{5.2}$$

where $\phi_i(y) = -f_i'(y)/f_i(y)$. Now $\bar{y}_i = E(y_i \mid R)$ may be obtained by taking expectations of each term in the series (5.2). We expand $\phi_i(m_i)$ about $\phi_i(\bar{y}_i)$ and, subject to certain regularity conditions, obtain

$$\bar{y}_i = \tfrac{1}{2}(\bar{y}_{i+1} + \bar{y}_{i-1}) - \tfrac{1}{2}\phi(\bar{y}_i)(\bar{y}_i - \bar{y}_{i-1})^2 + \mathcal{O}(N^{-3}) \tag{5.3}$$

or, to the same order,

$$\bar{y}_i \approx \tfrac{1}{2}(\bar{y}_{i+1} + \bar{y}_{i-1}) - \tfrac{1}{2}\phi_i(\bar{y}_i)(\bar{y}_i - \bar{y}_{i-1})(\bar{y}_{i+1} - \bar{y}_i) \tag{5.4}$$

so that

$$(\bar{y}_{i+1} - \bar{y}_i)^{-1} - (\bar{y}_i - \bar{y}_{i-1})^{-1} = -\phi_i(\bar{y}_i) + \mathcal{O}(N^{-1}). \tag{5.5}$$

We suggest using approximate scores $\{\tilde{y}_i\}$ which satisfy

$$(\tilde{y}_{i+1} - \tilde{y}_i)^{-1} - (\tilde{y}_i - \tilde{y}_{i-1})^{-1} = -\phi_i(\tilde{y}_i). \tag{5.6}$$

In general these approximations cannot be expected to hold for the extremes of the $\{\tilde{y}_i\}$. In the circumstances considered here, y is constrained to the positive half of the real line and $\phi_i(y)$ is finite and regular around $y=0$. In these circumstances, the approximation (5.5) holds for the smallest $\{\bar{y}_i\}$, taking $\bar{y}_0 = 0$. In the right tail, however, some modification of our approach is necessary.

We suppose that we are prepared to right-censor some proportion, ϵ, of the observations. That is, we assume simple type II censoring of the largest $N\epsilon = (N-k)$ observations. We may then derive an approximate expression for the spacing between $\bar{y}_k$ and $\bar{y}_{k-1}$. The approach is similar to that adopted above; we first condition upon y_{k-1} and show that

$$\begin{aligned} E(y_k - y_{k-1} \mid y_{k-1}, R) = {} & \left[\sum_{j \geq k} \lambda_j(y_{k-1})\right]^{-1} + \\ & + \left[\sum_{j \geq k} \lambda_j(y_{k-1})\right]^{-2} \lambda_k'(y_{k-1})/\lambda_k(y_{k-1}) + \\ & + \mathcal{O}(N^{-2}) \end{aligned} \tag{5.7}$$

where $\lambda_k(y) = f_k(y)/F_k(y)$. As before, we expand the right-hand side about $\bar{y}_k$ and obtain

$$\bar{y}_k - \bar{y}_{k-1} = \left[\sum_{j>k} \lambda_j(\bar{y}) + \phi_k(\bar{y}_k)\right]^{-1} + \mathcal{O}(N^{-2}). \tag{5.8}$$

The last spacing between our approximate scores is, therefore, given by

$$(\tilde{y}_k - \tilde{y}_{k+1})^{-1} = \sum_{j>k} \lambda_j(\tilde{y}_k) + \phi_k(\tilde{y}_k). \tag{5.9}$$

We may then re-express (5.6) and (5.9) as

$$(\tilde{y}_i - \tilde{y}_{i-1})^{-1} = \sum_{i \leqslant j \leqslant k} \phi_j(\tilde{y}_j) + \sum_{j>k} \lambda_j(\tilde{y}_k),\ i \leqslant k,\ = 0, \text{ otherwise.} \tag{5.10}$$

The argument may readily be extended to allow for progressive type II censoring. Finally, our approximate scores satisfy the following non-linear equations in the spacings,

$$(\tilde{y}_i - \tilde{y}_{i-1}) = \left[\sum_{j \geqslant i} \zeta_j(\tilde{y}_j)\right]^{-1},\ y_i \text{ uncensored},\ = 0, \text{ otherwise.} \tag{5.11}$$

where $\zeta_j(y)$ is, as in (2.4), either $\lambda_j(y)$ or $\phi_j(y)$ depending upon whether or not y is censored.

For the accelerated failure family considered here, $f_i(y)=\theta_i f(\theta_i y;\gamma)$ and $\zeta_i(y)=\theta_i\zeta(\theta_i y)$. Further, the density f represents a mixture of exponentials, and in these circumstances the expressions $\{\zeta(\theta_i\bar{y}_i)\}$ may be regarded as empirical Bayes estimates of the frailties $\{\xi_i\}$, for if the i-th individual is known to have failed at y,

$$E(\xi_i \mid Y_i = y;\gamma) = \phi(\theta_i y) = (1+\gamma)/(1+\theta_i y) \tag{5.12}$$

and, for right censored observations,

$$E(\xi_i \mid Y_i > y;\gamma) = \lambda(\theta_i y) = 1/(1+\theta_i y) \tag{5.13}$$

so that $\zeta(\theta_i\bar{y}_i)$ are approximate posterior expectations of $\{\xi_i\}$ given the rank vector, R.

The approximation of the matrix Σ, $\Sigma_{ij}=\mathrm{Cov}(y_i,y_j \mid R)$, presents more serious difficulties. Elsewhere (CLAYTON AND CUZICK [9]) we have suggested that, conditional upon R, the frailties $\{\xi_i\}$ are approximately independent. Writing $\{\Delta_i\}$ for the spacings $\{y_i - y_{i-1}\}$, we eventually obtain

$$\mathrm{Cov}(\Delta_i, \Delta_j \mid R) \approx \delta_{ij}\bar{\Delta}_i^2 - \bar{\Delta}_i^2\bar{\Delta}_j^2 \sum_{\substack{k \geqslant i \\ k \geqslant j}} \theta_k^2\, \zeta'(\theta_k\bar{y}_k) \tag{5.14}$$

where δ_{ij} is the Kroenecker delta. Note that $-\zeta'(\theta_i\bar{y}_i)$ approximates the posterior *variance* of ξ_i given the ranks, R.

Although this expression seems to give an adequate approximation, more careful examination suggests that it does not provide a consistent estimator of $N\Sigma$. Instead we note that conditional on R the random sequence $(y_i-\bar{y}_i)$, i=1,...,N, is a continuous-valued Markov chain. A normal approximation suggests that the variance-covariance matrix can be approximated by the inverse of a tridiagonal matrix. We obtain a similar prediction in the next section.

6. NON-PARAMETRIC M.L. ESTIMATION OF $\Lambda_0(t)$

The replacement of $y_i=\Lambda_0(t_i)$ by the scores $\bar{y}_i$ in the approximate E-step of the algorithm may be regarded simply as an attempt to construct a non-parametric estimator of the integrated baseline hazard function, $\Lambda_0(t)$. In this section we discuss an alternative approach which leads to an almost identical method.

BENNETT [4] suggested a semi-parametric approach to estimation in the proportional odds model ($\gamma=1$). In essence, his method amounts to a rather heuristic approach to non-parametric maximum likelihood estimation of $\Lambda_0(t)$ or, more accurately, of its values $\{y_i\}$, and derivatives at the observed failure times $\{t_i\}$, since the ML estimates of β are unaffected by the course of $\Lambda_0(t)$ between observed failure times. Although a more formal approach along the lines indicated by JOHANSEN [17] could be used, see Gill's discussion of CLAYTON AND CUZICK [9], in this section we reproduce Bennett's argument in the context of the wider Pareto family of rank regression models.

Bennett advocated simultaneous ML estimation of $\{y_i\}$ and β by Newton-Raphson iteration. For

comparison with our work, however, we shall consider maximization of his 'likelihood' by the two-step method described in section 3. We use quotes because his method is not strictly ML estimation. The maximum of the true likelihood over the space of all non-negative functions, $\lambda_i(t)$, is infinite, and the support for the function is concentrated onto the points, $\{t_i\}$, of observed failures. Bennett suggested avoidance of these difficulties by approximating the Jacobean term in (2.1) by

$$(\partial y / \partial t)_i = \lambda_0(t_i) \approx (y_i - y_{i-1}) / (t_i - t_{i-1}) \tag{6.1}$$

This is equivalent to modelling $\lambda_0(t)$ by a step function with discontinuities at the observed $\{t_i\}$. This argument is identical to that used by BRESLOW [6], [7] in proposing an estimator of $\Lambda_0(t)$ for the proportional hazards model ($\gamma=0$). For censored samples, the argument is modified by taking the discontinuities as occuring at the observed *uncensored* failure times. In the notation of (4.2) this involves replacing the subscripts i and $i-1$ in (6.1) by j_i and j_{i-1}.

To ease the notation we consider the uncensored case. The first step in the two-step recursion now involves estimation of the $\{y_i\}$ given that they are independently distributed with densities $\{\theta_i f(\theta_i y)\}$ where $f(y)$ is the standard Pareto density with parameter γ. Differentiation of Bennett's 'likelihood' yields

$$\partial \ell / \partial y_i = -\theta_i \phi(\theta_i y_i) + (y_i - y_{i-1})^{-1} - (y_{i+1} - y_i)^{-1} \tag{6.2}$$

and

$$\frac{\partial^2 \ell}{\partial y_i \partial y_j} = \begin{cases} (y_i - y_{i-1})^{-2} & , (j=i), \\ -\theta_i^2 \phi'(\theta_i y_i) - (y_i - y_{i-1})^{-2} - (y_{i+1} - y_i)^{-2} & , (j=i-1), \\ (y_{i+1} - y_i)^{-2} & , (j=i+1), \\ 0 & , \text{otherwise.} \end{cases} \tag{6.3}$$

The estimates of $\{y_i\}$ are obtained by solving the equations obtained by setting (6.2) to zero. These equations are identical to those for our approximate scores $\{\overline{y}_i\}$, (5.6), so that Bennett's method will lead to the same point estimates as the approximate maximum marginal likelihood method we have proposed. Thus, the difficulties implicit in Bennett's likelihood construction together with the fact that his method involves estimation of more parameters than data points seem to be of no consequence, at least as far as point estimation of β is concerned.

By considering the parallels between equations (3.2) and (4.5) we see that Bennett's method will likewise lead to correct asymptotic estimates of the variance of the estimates of β, *if* minus the inverse of the second derivative matrix (6.3) correctly estimates Σ, the variance-covariance matrix of the $\{y_i\}$ conditional upon the rank vector R. This remains to be proved. The matrix (6.3) is tridiagonal and may be economically inverted. Alternatively, we may use the Neumann lemma to obtain a series expansion for the inverse of (6.3), and the first two terms yield the variance estimator (5.14).

7. Other approaches and extensions.

Some general results related to semi-parametric regression have been recently set out by BEGUN, HALL, HUANG AND WELLNER [2]. Two other approaches more specifically directed at estimation in the semi-parametric proportional odds model have been published recently. The first of these, see PETTITT [26], applies the more general approach to rank regression problems proposed by PETTITT [25]. This, like our method, attempts to extend the marginal likelihood approach used by PRENTICE [29] in developing the censored linear rank test statistics to the problem of point and interval estimation of β. Pettitt's approach is to approximate the log of the marginal likelihood (4.1) resp. (4.2) by a Taylor expansion about $\beta=0$ up to quadratic terms in β. Essentially, the calculations then reduce to those for weighted least squares, the observed survival times being replaced by the scores given by Prentice.

Normally, our quasi-EM algorithm will start from $\beta=0$. The initial scores are the order statistics for i.i.d. Pareto variates and the first derivative of the log marginal likelihood gives the score test for

$\beta=0$. Pettitt's approach is equivalent to applying the correction (4.5) to obtain an approximation to the inverse of the second derivative matrix of the log marginal likelihood, and carrying out one step of the Newton-Raphson iteration for maximizing the likelihood.

In fact, this represents one step of an improved iterative method to compute our estimates. After this first step, we would go on to recompute new scores, $\{\bar{y}_i\}$, and the new value for the vector of first derivatives. It might not be necessary to refine our estimate of the second derivative matrix until convergence when the final variance estimate is required.

Given the possibility of continuing the iteration to obtain approximate MML estimates, there seems little to recommend acceptance of the first step estimates. They are not consistent and will only be acceptable when all coefficients in β are close to zero.

Another method has been proposed by McCullagh [21] for the proportional odds model. This generalizes methods orginally developed for ordered categorical data (McCullagh [20]). For a restricted number K of ordered categories, the unknown transformation g of (1.4) only enters as (K-1) unknown 'cut points' which in McCullagh's earlier work were simply estimated by maximum likelihood. The generalization to a full rank regression method invalidates this approach since the number of unknown parameters becomes more than the number of data points, and McCullagh proposes a quasi-likelihood (weighted least squares) approach based upon conditional moments of partial sums.

The regression method for ordered categorical data along the lines indicated by McCullagh represents a rank-regression problem with a particularly simple form of interval censoring. Observations are recorded only as lying in the intervals between adjacent 'cut points', $\{\theta_k\}$. While in a parametric problem $\{\theta_k\}$ are known, either the allowance for an arbitrary monotone transformation or deficiencies of recording means that frequently they must be regarded as unknowns. Eliminating these nuisance parameters from the problem by conditioning upon the marginal distribution of the categorical response is equivalent to basing inference upon a marginal likelihood for an interval censoring scheme rather similar to progressive type II right-censoring, i.e. a scheme in which the position of the cut-points $\{\theta_i\}$ in the ranking of the underlying observations is predetermined. The similarity between latent variable models for ordered categorical data and approaches based upon the marginal likelihood for tied ranked data has been discussed by Pettitt [27].

The approach we have used here is suitable for extension to this and more complex forms of censoring. The parametric kernel of the likelihood (4.2) can be modified to take account of the interval censoring, and the range of integration becomes the set of underlying vectors, β, consistent with the observed ordered categorical measurement. We must then devise alternative methods for computing the scores $\{\bar{y}_i\}$ for such schemes. Although such data are fairly commonplace in practice, the published methodology (apart from that using fully parametric methods) goes little further than non-parametric estimation of the survival function and simple 2-sample non-parametric tests (Peto [22]; Turnbull [30]; Peto et al [24].

The generalization of the heterogeneous frailty model (1.8) to encompass frailty shared by related individuals is discussed by Clayton and Cuzick [9]. Individuals are blocked within families or sibships $(S_1,\ldots,S_M)$ and different individuals within each sibship are assumed to share the same frailty. This extension requires modification of the simple parametric likelihood which forms the integrand of the marginal likelihood, and the scores become

$$\bar{y}_i = E(y_i \mid R, S_1, \ldots, S_M; \beta, \gamma). \tag{7.1}$$

Approximate methods for computing these scores are given by Clayton and Cuzick [9] and may be justified rather more formally along the lines indicated in section 5 or 6. Again they may be expressed in terms of approximate empirical Bayes estimates of the (shared) frailties.

Finally, the method may be further generalized by considering individuals as drawn from different strata in which different baseline hazard functions, $\lambda_0(t)$, apply. Now, a different monotone transformation must be used in each stratum to meet the model assumptions and we must now base inference upon a set of rank vectors, one for each stratum. This generalization follows closely the corresponding generalization for parametric likelihood in the Cox model, see Kalbfleisch and Prentice [19],

section 4.4. In the special case where there are two strata and each family includes one member from each stratum, we have the father-son problem discussed by CLAYTON [8] and CUZICK [13]. The treatment of that problem along the present lines is also given by CLAYTON AND CUZICK [9].

8. NUMERICAL RESULTS.

In our earlier paper (CLAYTON AND CUZICK [9]) we evaluated our method for estimation of the parameter γ in the father-son association problem using Monte Carlo simulation. Here we evaluate our method for estimation of β for known γ for the simpler regression problem. We consider the case of a single covariate, z, taking values 0 for half the observations and 1 for the remaining half. We generated four sets of 1000 data-sets, all with $\beta=1$; the first two sets were of size N=20, and the second two sets were of size N=50. For each sample size, we considered the cases $\gamma=1$ and $\gamma=2$.

Table 1

Data-set	$\gamma=1$		$\gamma=2$	
	Mean (Median)	S.D. (Mean Est*)	Mean (Median)	S.D. (Mean Est*)
N=20				
k=0	1.24 (1.18)	0.845 (0.814)	1.22 (1.18)	1.07 (1.03)
k=2	1.24 (1.18)	0.850 (0.813)	1.22 (1.18)	1.07 (1.02)
k=5	1.21 (1.14)	0.864 (0.829)	1.21 (1.17)	1.08 (1.03)
N=50				
k=0	1.06 (1.05)	0.490 (0.506)	1.05 (1.05)	0.631 (0.645)
k=5	1.06 (1.05)	0.492 (0.507)	1.05 (1.05)	0.631 (0.645)

**The mean of the 1000 estimates of the standard deviation of $\hat{\beta}$ using (5.14).*

We analyzed these data under varying degrees of simple type II censoring; for N=20 we considered censoring of the largest k=0, 2 or 5 observations, while for N=50 we considered k=0 or 5. Table 1 shows the mean, median and standard deviation of the estimates of β in each series of simulations, together with the mean of the estimates of the standard error.

For the smaller sample size, the estimate was substantially biased but for N=50 the method performed well (at the time of writing it is not known whether or not the equivalent fully parametric method behaves similarly). Furthermore, although artificially added right censoring of up to 10% of the observations is of scarcely any consequence (see section 5), it appears that such censoring is not necessary in practice.

REFERENCES

1. R.J.BAKER AND J.A.NELDER (1978). *The GLIM System, Release 3.* , Oxford, Numerical Algorithms Group.
2. J.M.BEGUN, W.J.HALL, W.M.HUANG AND J.A.WELLNER (1983). Information and asymptotic efficiency in parametric and non-parametric models, *Ann.Statist. 11* ,432-452.
3. S.BENNETT (1983a). Log-logistic regression models for survival data. *Appl.Statist, 32* ,165-171.
4. S.BENNETT (1983b). Analysis of survival data by the proportional odds model. *Statist. in Medicine, 2* , 273-277.
5. G.E.P.BOX AND D.R.COX (1964). An analysis of transformations. *J.R.Statist.Soc.B, 26* , 211-252.
6. N.E. BRESLOW (1972). Contribution to the discussion of the paper by D.R. Cox, Regression Models and Life Table, *J.R.Statist.Soc.B, 34* , 216-217.
7. N.E. BRESLOW (1974). Covariance analysis of censored survival data, *Biometrics 30* , 89-100.
8. D.G.CLAYTON (1978). A model for association in bivariate life-tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika, 65* , 141-151.
9. D.G.CLAYTON AND J.CUZICK (1985a). Multivariate generalisations of the proportional hazards model (with Discussion). *J.R.Statist.Soc.A. 148*
10. D.G.CLAYTON AND J.CUZICK (1985b). The EM algorithm for Cox's regression model using GLIM. Preprint.
11. D.R.COX (1972). Regression models and life-tables (with Discussion). *J.R.Statist.Soc.B, 34* ,187-220.
12. D.R.COX (1975). Partial likelihood. *Biometrika, 62* , 269-276.
13. J.CUZICK (1982). Rank tests for association with right-censored data. *Biometrika, 69* , 351-364.
14. J.CUZICK (1985). Asymptotic properties of censored linear rank tests. *Ann.Statist, 13* , 133-141.
15. A.P.DEMPSTER, N.M.LAIRD AND D.B.RUBIN (1977). Maximum likelihood from incomplete data via the EM algorithm. *J.R.Statist.Soc.B, 39* , 1-38.
16. D.P.HARRINGTON AND T.R.FLEMING (1982). A class of rank test procedures for censored survival data. *Biometrika, 69* , 553-566.
17. S.JOHANSEN (1983). An extension of Cox's regression model. *Int.Statist.Rev. 1* , 165-174.
18. J.D.KALBFLEISCH AND R.L.PRENTICE (1973). Marginal likelihoods based upon Cox's regression and life model. *Biometrika, 60* , 267-278.
19. J.D.KALBFLEISCH AND R.L.PRENTICE (1980). *The Statistical Analysis of Failure Time Data.* , New York: Wiley.
20. P.MCCULLAGH (1980). Regression models for ordinal data. *J.R.Statist.Soc.B, 42* , 109-142.
21. P.MCCULLAGH (1984). On the elimination of nuisance parameters in the proportional odds model. *J.R.Statist.Soc.B, 46* , 250-256.
22. R.PETO (1973). Experimental survival curves for interval-censored data. *Appl.Statist. 22* , 86-91.
23. R.PETO AND J.PETO (1972). Asymptotically efficient rank invariant test procedures (with discussion). *J.R.Statist.Soc.A, 135* , 185-206.
24. R.PETO, M.C.PIKE, N.E.DAY, R.G.GRAY, P.N.LEE, S.PARISH, J.PETO, S.RICHARDS AND J.WAHRENDORF (1980). Guidelines for simple, sensitive significance tests for carcinogenic effects in log-term animal experiments. In *Long-term and short-term screening assays for carcinogens: A critical appraisal.* Lyon: IARC.
25. A.N.PETTITT (1983). Approximate methods using ranks for regression with censored data. *Biometrika, 70* , 121-132.
26. A.N.PETTITT (1984a). Proportional odds models for survival data and estimates using ranks. *Appl.Statist. 33* , 169-175.
27. A.N.PETTITT (1984b). Tied, grouped continuous and ordered caregorical data: a comparison of two models. *Biometrika, 71* , 35-42.
28. R.L.PRENTICE (1973). Exponential survivals with censoring and explanatory variables. *Biometrika, 60* , 279-288.
29. R.L.PRENTICE (1978). Linear rank tests with right censored data. *Biometrika, 65* , 153-158.

30. B.W.TURNBULL (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J.R.Statist.Soc.B, 38* , 169-173.
31. J.W.VAUPEL, K.G.MANTON AND E.STALLARD (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography, 16* , 439-454.

On asymptotic inference about intensity parameters of a counting process.

Kacha Dzhaparidze
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

The Cox regression model may be viewed as a special case of the general model described in this paper via the pair $(\bar{A}_t, \Psi_t)$ of predictable characteristics of an r-variate counting process $\mathbb{N}_t = (N_t^1, ..., N_t^r)$, associated with its compensator $\mathbb{A}_t = (A_t^1, ..., A_t^r)$ as follows: $\bar{A}_t = A_t^1 + ... + A_t^r$ and $\Psi_t = d\mathbb{A}_t / d\bar{A}_t$. It is supposed that the latter characteristic involves the real valued parameter β, i.e. $\Psi_t = \Psi_t(\beta)$, to be estimated by means of a given sample path of $\{\mathbb{N}_t,\ 0 \leqslant t \leqslant 1\}$. Treating this problem in its asymptotic setting, we consider our experiment as n-th in a sequence of experiments, and let $\bar{A}_t$ meet Condition I of asymptotic stability. Under this and certain additional conditions introduced on demand, we study asymptotic properties of the estimator $\hat{\beta}$ for β, which is in fact the Cox estimator extended to our situation. In particular, we characterize the consistency and asymptotic normality of $\hat{\beta}$ by estimating the probability of large deviations, and then showing the convergence in all moments of the distribution of $\hat{\beta}$ to a normal law. Finally, it is shown that $\hat{\beta}$ is the best within a class of (*regular*) estimators in the sense that none of them can have an asymptotic distribution that is less spread out than that of $\hat{\beta}$.

1. Introduction.

1.1. Statistical inference on counting processes attracts considerable attention in the literature of recent years; see the bibliography where a number of related references is given which may serve as a source for many further references. Typically, the approach taken in these works is inspired by developments in the theory of stochastic processes related to the notion of martingales, see e.g. Shiryayev [23], as well as by developments in the asymptotic theory of statistical decisions, see e.g. Le Cam [17] or Ibragimov and Has'minskii [11]; also Greenwood and Shiryayev [9].

Within the framework of the theory of stochastic processes, these processes are defined on a complete probability space $(\Omega, \mathcal{F}, P)$ equipped with a nondecreasing family $\{\mathcal{F}_t, t \geqslant 0\}$ of right-continuous sub-σ-algebras of $\mathcal{F}$ augmented by sets from $\mathcal{F}$ of zero probability. For the sake of simplicity, we discuss only the case in which $t \in [0,1]$.

Let $\mathbb{N} = (\mathbb{N}_t, \mathcal{F}_t, P)$ be an r-variate counting process which by definition consists of components $N^i, i = 1, ..., r$ whose sample paths are step functions: $N_0^i = 0$, $N_t^i - N_{t-}^i = \Delta N_t^i = 0$ or 1, $\Delta N_t^i \Delta N_t^j = 0$ if $i \neq j$ (no two component processes jump at the same time), and $N_1^i < \infty$ P - a.s. With $\mathbb{N}$ one may associate an r-variate predictable increasing process $\mathbb{A} = \{\mathbb{A}_t, \mathcal{F}_t, P\}$ such that $\mathbb{N} - \mathbb{A} \equiv \mathbb{M} = \{\mathbb{M}_t, \mathcal{F}_t, P\}$ is an r-variate local square integrable martingale with the predictable quadratic characteristic $<\mathbb{M}>_t = \mathrm{diag}\mathbb{A}_t - [\mathbb{A}]_t$ (see Lemma 3.1).

If, in addition, the filtration is of special form $\mathcal{F}_t = \sigma\{\omega : \mathbb{N}_s, s \leqslant t\}$ then the probability measure P is completely defined by the compensator $\mathbb{A}$ (in the sense of Liptser and Shiryayev [18], Section 18.3). Hence in this case the statistical model for the observed phenomena may be completely specified in terms of the compensator $\mathbb{A}_t$ or, for convenience, in terms of the so-called $(P, \mathcal{F}_t)$- predictable characteristics $(\bar{A}_t, \Psi_t)$ of $\mathbb{N}_t$, associated with $\mathbb{A}_t$ by the following relations $\Psi_t = d\mathbb{A}_t / d\bar{A}_t$ and $\bar{A}_t = \mathbb{1}_r^T \mathbb{A}_t$ (here $\mathbb{1}_r = \mathrm{col}(1, ..., 1)$, and T indicates a transposition). Obviously, the first of these characteristics is the compensator of $\bar{N}_t = N_t^1 + ... + N_t^r$, while the r-variate nonnegative predictable process $\Psi = \{\Psi_t, \mathcal{F}_t, P\}$ consists of components $\Psi_t^i, i = 1, ..., r$, Ψ_t^i being, roughly speaking, the probability of having a jump of N_t^i at time t, given $\mathcal{F}_{t-}$ and given that $\bar{N}_t$ jumps at time t; see Brémaud [4], pp. 34

and 236.

1.2. In applications the latter characteristic is usually parametrized: it is restricted to a certain parametric family $\Psi\in\{\Psi(\beta),\beta\in\mathfrak{B}\}$ of nonnegative $\mathfrak{F}_t$-predictable processes for each admissable value of the parameter $\beta\in\mathfrak{B}$.

In such a case β is 'the parameter of interest' - inference about β is required by means of a given sample path of $\{\mathbb{N}_t,0\leqslant t\leqslant 1\}$ drawn according to the pair $(\bar{A}_t,\Psi_t)_\beta$ of the characteristics of $\mathbb{N}$ for an unknown β and, typically, for the characteristic $\bar{A}_t$ specified only up to the restrictions of a general nature (to be introduced below). Actually, $\bar{A}_t$ itself may depend on the parameter of interest β, as well as on certain nuisance quantities, as it is illustrated by the following:

EXAMPLE 1.1. Let $\{P_{\alpha,\beta},\alpha\in\mathfrak{A},\ \beta\in\mathfrak{B}\}$ be a family of probability measures, where $\mathfrak{A}$ is a set of deterministic nonnegative and nondecreasing functions $\alpha=\alpha_t$, $0\leqslant t\leqslant 1$, and $\mathfrak{B}$ an open set of R^1. For each $\alpha\in\mathfrak{A}$ and $\beta\in\mathfrak{B}$ let $\mathbb{N}=(\mathbb{N}_t,\mathfrak{F}_t,P_{\alpha,\beta})$ be an r-variate counting process of the Poisson type (LIPTSER AND SHIRYAYEV [18], p. 249) defined on the stochastic basis $(\Omega,\mathfrak{F},\{\mathfrak{F}_t,\ 0\leqslant t\leqslant 1\},P_{\alpha,\beta})$, with the compensator of form

$$\mathbb{A}_t=\mathbb{A}_t(\alpha,\beta)=\int_0^t\Phi_s(\beta)d\alpha_s,\ 0\leqslant t\leqslant 1 \tag{1.1}$$

where $\Phi(\beta)$ is an r-variate nonnegative $\mathfrak{F}_t$-predictable process for each $\beta\in\mathfrak{B}$. Obviously, the pair of the $(P_{\alpha,\beta},\mathfrak{F}_t)$- predictable characteristics of the process $\mathbb{N}$ is given by the following relations

$$\Psi_t(\beta)=\Phi_t(\beta)/\bar{\Phi}_t(\beta)\quad\text{with}\quad\bar{\Phi}_t(\beta)=\mathbb{I}_r^T\Phi_t(\beta),\ \text{and}\ \bar{A}_t(\alpha,\beta)=\int_0^t\bar{\Phi}_s(\beta)d\alpha_s \tag{1.2}$$

The most popular special case of the Cox regression model for censored survival data specifies these characteristics as follows:

$$\Psi_t^i(\beta)=Y_t^ie^{\beta Z_t^i}/\sum_{i=1}^r Y_t^ie^{\beta Z_t^i},\ \bar{A}_t(\alpha,\beta)=\int_0^t\sum_{i=1}^r Y_s^ie^{\beta Z_s^i}d\alpha_s \tag{1.3}$$

with certain $\mathfrak{F}_t$-predictable processes Y_t^i and Z_t^i, free from β; see, e.g. ANDERSEN AND GILL [1] (or, for a little more general model, PRENTICE AND SELF [22]). These authors and later BEGUN et al. [2] have shown that under wide conditions the particular estimator $\hat{\beta}$ for β, defined by the relation

$$\sup_{\beta\in\mathfrak{B}}\int_0^1\ln^T\Psi_s(\beta)d\mathbb{N}_s=\int_0^1\ln^T\Psi_s(\hat{\beta})d\mathbb{N}_s,\quad\ln\Psi=\operatorname{col}\{\ln\Psi^i,i=1,\ldots,r\} \tag{1.4}$$

possesses the desired asymptotic properties (to be specified in the next section)

Obviously, if $\Psi_s(\beta)$ is a sufficiently smooth function of β, then the estimator $\hat{\beta}$ of β is well defined by condition (1.4) also for the general set up discussed at the beginning of this subsection (and not only for the special Cox model; see (1.3)). Naturally, one can expect that under circumstances similar to those of the papers mentioned above, the estimator $\hat{\beta}$ preserves its desired properties. In the present paper this conjecture is confirmed, furthermore, a refined characterization of these properties is given (cf. EFRON [7]).

Note that unlike ANDERSEN AND GILL [1] here only inference on the real valued parameter β is discussed, while the abstract parameter α in (1.3) (or (1.2)) is considered as a nuisance quantity.

2. Asymptotic Inference.

2.1. Following the usual device of asymtotic theory (Le Cam [17], Ibragimov and Has'minskii [11]), we suppose that what is observed is an outcome of the experiment

$$\mathcal{E}_n = (\Omega^n, \mathcal{F}^n, \{\mathcal{F}_t^n, 0 \leqslant t \leqslant 1\}, \{P^n\}) \tag{2.1}$$

(with a certain family of probability measures $\{P^n\}$), which is actually n-th in the sequence of experiments $\mathcal{E}_1, \mathcal{E}_2, \ldots$. Fix P^n at the right-hand side of (2.1), and define on that stochastic basis an r_n-variate counting process $\mathbb{N}^n = \{\mathbb{N}_t^n, \mathcal{F}_t^n, P^n\}$ where $r_n, n = 1,2,\ldots$ is a nondecreasing sequence of integers. As above, to the compensator $\mathbb{A}^n = \{\mathbb{A}_t^n, \mathcal{F}_t^n, P^n\}$ of $\mathbb{N}^n$ relate the pair $(\bar{A}_t^n, \Psi_t^n)$ of the $(P^n, \mathcal{F}_t^n)$-predictable characteristics, and let it depend on $\beta \in \mathcal{B}$ in the way described at the beginning of Subsection 1.2 (recall that $\bar{A}_t^n$ is not fully specified - in Example 1.1, for instance, it depends also on a nuisance parameter α which however is suppressed from the notation here and below).

The class of all admissible pairs $(\bar{A}_t^n, \Psi_t^n)_{\beta \in \mathcal{B}}$ of the $(\{P_\beta^n, \beta \in \mathcal{B}\}, \mathcal{F}_t^n)$ -predictable characteristics of $\mathbb{N}^n$ determines the family of the probability measures $\{P^n\}$ in (2.1). The following basic condition restricts this class by an asymptotic stability requirement on the sequence[1] $F_t^n \equiv \bar{A}_t^n / k_n = \mathbb{I}_{r_n}^T \mathbb{A}_t^n / k_n$ for some unboundedly increasing sequence of numbers k_n, $n = 1,2,\ldots$.

Condition I. For each admissible pair $(\bar{A}_t^n, \Psi_t^n)_\beta$ of the $(P_\beta^n, \mathcal{F}_t^n)$-predictable characteristics there exists a continuous deterministic non-decreasing function F_t such that $F_t^n \to F_t$ in P_β^n probability as $n \to \infty$, each $t, 0 \leqslant t \leqslant 1$.

Remark 2.1. In fact, by lemma 1 of McLeish [20], p. 146 the continuity of F_t implies $\sup_{0 \leqslant t \leqslant 1} |F_t^n - F_t| \to 0$ in P_β^n probability as $n \to \infty$, for each $\beta \in \mathcal{B}$.

Define now the estimator $\hat{\beta}_n$ for β by condition (1.4) with $\mathbb{N} = \mathbb{N}^n$ and $\Psi = \Psi^n$. On deriving asymptotic (as $n \to \infty$) properties of $\hat{\beta}_n$, we require some regularity conditions on $\Psi^n(\beta)$; see Conditions II-IV in Section 5.

Condition II requires differentiability (in a certain sense) of $\sqrt{\Psi^n(\beta)} = \operatorname{col}\{\sqrt{\Psi^{in}}, i = 1,\ldots,r_n\}$ and existance of a positive number v - the limit of $\int_0^1 |(\partial/\partial\beta)\sqrt{\Psi^n(\beta)}|^2 \mathrm{d}F^n$ in P_β^n probability as $n \to \infty$. Condition III (of the Lindeberg type), together with Condition II, leads to the conclusion of Corollary 5.1 needed for deriving asymptotic normality $N(0, 1/4v)$ of the estimator $\hat{\beta}_n$.

As for Condition IV, it permits us (via Lemma 5.1) to apply a generalized version of Theorem 5.1 of Ibragimov and Has'minskii [11], (Section 1.5: Inequalities for Probabilities of Large Deviations) due to Sieders and Dzhaparidze [24], the conclusion of which can be informally described as follows: Let an estimator $\hat{\beta}_n$ for β be defined by maximizing with respect to β a certain functional of observations (e.g. the likelihood function). If this functional satisfies certain conditions, similar to the conditions imposed on the likelihood function in the above mentioned Theorem 5.1, then the estimator $\hat{\beta}_n$ is not only consistent (in P_β^n probability), but also the following holds: for sufficiently large values of n the P_β^n-probability that $2\sqrt{k_n v}(\hat{\beta}_n - \beta)$ exceeds in absolute value a (sufficiently large) number H is less then $C_0 \exp(-c_0 H^2)$, with some positive constants c_0 and C_0.

Hence, this way we get the first main result of Section 5 - the refinement of consistency of the estimator $\hat{\beta}_n$ (Proposition 5.1).

The second main result concerns the refinement of asymptotic normality of $\hat{\beta}_n$ based on a generalization of Theorem 10.1 of Ibragimov and Has'minskii [11], (p. 103): if the generalized version of Theorem 5.1 holds (Sieders and Dzhaparidze [24]), as well as Corollary 5.1 and Lemma 5.2, then

1. For simplicity we do not index F^n and F by β on which they may actually depend; cf. Example 1.1.

all moments of $2\sqrt{k_n v}(\hat{\beta}_n - \beta)$ converge to the corresponding moments of the standard normal distribution (Proposition 5.2).

In the conclusion of Section 5, the results mentioned above are applied to the situation described in Example 1.1, specifically, to the Cox regression model (see (1.3)).

2.2. In discussing optimality properties of the estimator $\hat{\beta}_n$ in Section 6, we restrict our considerations to processes of the Poisson type; see Example 1.1 in which all the introduced quantities are indexed now by n, except the parameters α and β, of course.

In the first place we show the LAN property of the family $\{P^n_{\alpha,\beta}, \alpha \in \mathcal{A}, \beta \in \mathcal{B}\}$ of the probability measures defined on $\{\Omega^n, \mathcal{F}^n\}$; see Definition 5.1. Along with $P^n = P^n_{\alpha,\beta}$, let the probability measure $\underline{P}^n = P^n_{\underline{\alpha}^n,\underline{\beta}^n}$ be defined on $(\Omega^n, \mathcal{F}^n)$, where $\underline{\beta}^n = \beta + b/\sqrt{k_n} \in \mathcal{B}$, $b \in R^1$ and $\underline{\alpha}^n \in \mathbf{A}$ is defined by the relation $\sqrt{d\underline{\alpha}^n_t / d\alpha_t} = 1 + a_t/\sqrt{k_n}$, $a_t \in L^2(dF)$ with $F_t = F_t(\alpha,\beta)$ of Condition VI.2. Then $P^n \ll \underline{P}^n$, and $d\underline{P}^n/dP^n$ is given by (6.1). The above mentioned LAN property is stated in Proposition 6.1 which tells us that under the Conditions ΦI-ΦIII the logarithm of $d\underline{P}^n/dP^n$ is in fact asymptotically quadratic with the asymptotically normal linear term $\delta^n_{\alpha,\beta}(a,b)$, and the quadratic term $-\frac{1}{2}g_{\alpha,\beta}(a,b)$ where $g_{\alpha,\beta}(a,b)$ is the limit in P^n probability of the quadratic characteristic of $\delta^n_{\alpha,\beta}(a,b)$.

Condition ΦI.1 requires continuous differentiability of $\sqrt{\Phi^n(\beta)}$ (in certain sense), and Condition ΦI.2, together with Condition ΦII, determines the form of $g_{\alpha,\beta}(a,b)$. Condition ΦIII (of the Lindeberg type) ensures the asymptotic normality of the linear term $\delta^n_{\alpha,\beta}(a,b)$.

Having the LAN property, one can take advantage of its fairly general implications due to Le Cam and Hájek (see, e.g., IBRAGIMOV AND HAS'MINSKII [11], Ch. II and III, or MILLAR [21]). Specifically, our conclusions about asymptotic optimality properties of the estimator $\hat{\beta}_n$ are based on the application of Hájek's convolution theorem to the situation under consideration (see Theorem 6.1).

For these purposes, define first the class of regular estimators $\{\beta^n_R\}$ for β. Under the conditions ensuring the LAN property of the family $\{P^n_{\alpha,\beta}, \alpha \in \mathcal{A}, \beta \in \mathcal{B}\}$, at 'point' $\alpha \in \mathcal{A}, \beta \in \mathcal{B}$ (Proposition 6.1), the estimator β^n_R is called *regular* (at the point $\alpha \in \mathcal{A}, \beta \in \mathcal{B}$) if for some nondegenerate distribution function G^0_R the following weak convergence takes place:

$$\mathcal{L}\{\sqrt{k_n}(\beta^n_R - \underline{\beta}^n) \mid \underline{P}^n\} \Rightarrow G^0_R \tag{2.2}$$

uniformly for each $|b| < c$ whatever $c > 0$, and each bounded $a \in L^2(dF)$ ($\underline{\alpha}^n$, $\underline{\beta}^n$ and $\underline{P}^n$ being defined as above).

Now, Hájek's convolution theorem (BEGUN et al. [2]) tells us that G^0_R at the right-hand side of (2.2) can be represented as the convolution of a certain normal law with another distribution law, G^1_R say. By Proposition 6.2, in our special case $G^0_R = N(0, 1/4v) * G^1_R$ where $N(0, 1/4v)$ coincides with the asymptotic distribution of $\sqrt{k_n}(\hat{\beta}_n - \beta)$; see the previous subsection.

Since convolution 'spreads out mass', no regular estimator β^n_R can have an asymptotic distribution that is less spread out then $N(0, 1/4v)$. Thus, in this sense the estimator $\hat{\beta}_n$ (which is regular under the conditions of the previous subsection; see Theorem 6.1) is best within the class $\{\beta^n_R\}$.

The proof of the results just mentioned uses the fact that the neighborhood about α that shrinks at rate $k_n^{-1/2}$ in the directions $\{a\}$, defined above, is 'sufficiently rich' to include the function $(\partial/\partial\beta)\sqrt{\phi_t(\beta)} / \sqrt{\phi_t(\beta)}$ where $\phi_t(\beta)$ is the bounded limit in $P^n_{\alpha,\beta}$ probability of $\bar{\Phi}^n_t(\beta)/k_n$ (Condition ΦII). Simply, the set $\{a\}$ includes $(\partial/\partial\beta)\sqrt{\phi_t(\beta)} / \sqrt{\phi_t(\beta)}$; see Proposition 6.2.

2.3. This inclusion typically fails in situations in which $\{a\}$ is a low dimensional subspace of $L^2(dF)$, namely, in the frequently encountered situations in which 'the cumulative hazard function' α_t is also parametrized up to a certain number of nuisance parameters, and hence $\{a\}$ is taken as a linear subspace, $\mathbf{A} = \mathbf{A}(\alpha)$ say, spanned by the logarithmic derivatives of the density of α_t with respect to the nuisance parameters; see, e.g. EFRON [7], JARUPSKIN [12], BORGAN [3], HJORT [10]. According to these works the following conclusions can be drawn about the maximum likelihood estimator β^n_{ML} for β, defined by maximizing the likelihood function (see (6.1)) simultaneously with respect to the

parameter of interest β and the nuisance parameters.

Under certain regularity conditions $\mathcal{L}\{\sqrt{k_n}(\beta^n_{ML}-\beta^n)|P^n\}\Rightarrow N(0,\mathcal{I}^{-1}_{\alpha,\beta})$ with $\mathcal{I}_{\alpha,\beta}$ defined as in (6.3), and this means that no $R^1\times\mathbf{A}$-*regular* estimator β^n_{RA} can have an asymptotic distribution less spread then that of β^n_{ML}. In fact, the estimator β^n_{RA} is called $R^1\times\mathbf{A}$-*regular* if for some nondegenerate distribution function G^0_{RA} $\mathcal{L}\{\sqrt{k_n}(\beta^n_{RA}-\beta^n)|P^n\}\Rightarrow G^0_{RA}$ for each $b\in R^1, a\in\mathbf{A}$, whereas Proposition 6.2 tells us that G^0_{RA} may be represented as the convolution (6.3). In particular, β^n_{ML} is less dispersed then the estimator $\hat{\beta}_n$, for comparing their variances we have $\mathcal{I}^{-1}\leqslant(4v)^{-1}$ with equality iff $(\partial/\partial\beta)\sqrt{\phi_t(\beta)}/\sqrt{\phi_t(\beta)}\in\mathbf{A}$ (see Remark 6.3).

It is important, however, that there is a subclass of estimators for β within which no estimator has a less spread asymptotic distribution then $\hat{\beta}_n$ defined by (1.4). This is the subclass $\{\beta^n_R\}\subset\{\beta^n_{RA}\}$ of regular estimators defined as in the previous subsection by the condition: whatever the (bounded) direction $a\in L^2(dF)$ of approach to α there is some nondegenerate distribution function G^0_R such that (2.2) takes place. Of course, β^n_{ML} is not regular in this sense, as for $a\notin A$ a bias appears in its limiting distribution. However the estimator $\hat{\beta}_n$ is regular, and it is the best among $\{\beta^n_R\}$ since by Proposition 6.3 G^0_R may be always represented as the convolution $G^0_R=N(0,1/4v)*G^1_R$ (Theorem 6.1).

3. The likelihood ratio for counting processes

3.1. Let $(\Omega, \mathcal{F}, P)$ be a complete probability space with a filtration $\{\mathcal{F}_t,0\leqslant t\leqslant 1\}$ satisfying the usual conditions. Let $\mathbb{N}=\{\mathbb{N}_t,\mathcal{F}_t,P;0\leqslant t\leqslant 1\}$ be a multivariate (r-variate) counting process: $\mathbb{N}=\mathrm{col}\{N^1,\ldots,N^r\}$. Consider its Doob-Meyer decomposition $\mathbb{N}=\mathbb{M}+\mathbb{A}$ where $\mathbb{M}=\{\mathbb{M}_t,\mathcal{F}_t,P;0\leqslant t\leqslant 1\}$ is a local square integrable martingale, and $\mathbb{A}=\{\mathbb{A}_t,\mathcal{F}_t,P;0\leqslant t\leqslant 1\}$ a predictable compensator.

Lemma 3.1. *The quadratic variation and quadratic characteristic of* $\mathbb{M}$ *are given by the following relations:*

1) $[\mathbb{M}]=\mathrm{diag}\mathbb{N}-[\mathbb{A}]-[\mathbb{M},\mathbb{A}]-[\mathbb{A},\mathbb{M}]$

2) $\langle\mathbb{M}\rangle=\mathrm{diag}\mathbb{A}-[\mathbb{A}]$

Proof. By definition $[\mathbb{N}]=\mathrm{diag}\mathbb{N}$, and this gives 1). To get 2) take the compensator of both sides of 1).

□

Remark 3.1. Denote $\bar{N}=N^1+\ldots+N^r, \bar{N}=\bar{M}+\bar{A}$. From 2) it follows that

$$\langle\bar{M}\rangle_t=\bar{A}_t-[\bar{A}]_t=\int_0^t(1-\Delta\bar{A})d\bar{A},\quad \Delta\langle\bar{M}\rangle=(1-\Delta\bar{A})\Delta\bar{A}\ ,$$

hence $0\leqslant\Delta\bar{A}\leqslant 1$. For simplicity assume $\Delta\bar{A}<1$ (in fact one can easily dispence with this restriction; see e.g. Kabanov et al. [13] or [16]).

Remark 3.2. Consider $V_t=I_r-\Delta\mathbb{A}_t\otimes\mathbb{I}_r$ and $V_t^{-1}=I_r+(1-\Delta\bar{A}_t)^{-1}\Delta\mathbb{A}_t\otimes\mathbb{I}_r$ with $\mathbb{I}_r=\mathrm{col}\{1,\ldots,1\}$ and $I_r=\mathrm{diag}\mathbb{I}_r$. Then

$$\begin{aligned}\langle\mathbb{M}\rangle_t&=\int_0^t V\mathrm{diag}d\mathbb{A}\\&=\int_0^t\mathrm{diag}d\mathbb{A}\,V^T\\&=\int_0^t V^{1/2}\mathrm{diag}d\mathbb{A}\,V^{1/2T}\end{aligned}$$

with

$$V^{1/2} = I_r - (1-\sqrt{1-\Delta\bar{A}})\Delta\mathbb{A}/\Delta\bar{A} \otimes \mathbb{I}_r \cdot I(\Delta\bar{A}>0)$$

(satisfying $(V^{1/2})^2 = V$, of course), and

$$\begin{aligned}\operatorname{diag}\mathbb{A}_t &= \int_0^t V^{-1} d\langle\mathbb{M}\rangle \\ &= \int_0^t d\langle\mathbb{M}\rangle V^{-1T} \\ &= \int_0^t V^{-1/2} d\langle\mathbb{M}\rangle V^{-1/2T}\end{aligned}$$

with

$$V^{-1/2} = I_r + \frac{1-\sqrt{1-\Delta\bar{A}}}{\sqrt{1-\Delta\bar{A}}} \frac{\Delta\mathbb{A}}{\Delta\bar{A}} \otimes \mathbb{I}_r \cdot I(\Delta\bar{A}>0)$$

LEMMA 3.2. *Let* $\mathbb{Q}_t = \int_0^t V^{-1} d\mathbb{A} = \int_0^t (1-\Delta\bar{A})^{-1} d\mathbb{A}$ *and* $\mathfrak{M}_t = \int_0^t V^{-1} d\mathbb{M} = \mathbb{M}_t + [\mathbb{Q},\bar{M}]_t = \mathbb{M}_t + [\mathbb{A},\bar{\mathfrak{M}}]_t$ *where* $\bar{\mathfrak{M}}$ *is the sum of the component of* $\mathfrak{M}$. *Then*

1) $[\mathfrak{M}]_t = \operatorname{diag}\mathbb{N}_t + \int_0^t (1-\Delta\bar{N}) d[\mathbb{Q}]$,

2) $\langle\mathfrak{M}\rangle = \operatorname{diag}\mathbb{A} + [\mathbb{Q},\mathbb{A}]$.

PROOF. As $\Delta\mathbb{N}^{\otimes 2} = \operatorname{diag}\Delta\mathbb{N}$, $(1-\Delta\bar{N})^2 = (1-\Delta\bar{N})$ and $\Delta\mathbb{N}(1-\Delta\bar{N}) = 0$, 1) follows from

$$\Delta\mathfrak{M} = \Delta\mathbb{M} + \Delta\mathbb{Q}\Delta\bar{M} = \Delta\mathbb{N} - \Delta\mathbb{Q}(1-\Delta\bar{N}). \tag{3.1}$$

To get 2) take the compensators of both sides of 1).

□

3.2. Suppose that a probability measure P in addition to the probability measure P is given on a measurable space $(\Omega,\mathcal{F})$ with a filtration of special form $\mathcal{F}_t = \sigma\{\mathbb{N}_s : s \leqslant t\}, 0 \leqslant t \leqslant 1$. Along with $\mathbb{N} = (\mathbb{N}_t, \mathcal{F}_t, P)$, consider the counting process $\underline{\mathbb{N}} = (\mathbb{N}_t, \mathcal{F}_t, \underline{P})$ with compensator $\underline{\mathbb{A}} = (\underline{\mathbb{A}}_t, \mathcal{F}_t, \underline{P})$.

THEOREM 3.1. (see KABANOV et al. [15]).
1) For absolute continuity of $\underline{P}$ *with respect to* $P(\underline{P} \ll P)$ *the following conditions are necessary and sufficient:* P*-a.s.*

I. $\Delta\underline{\bar{A}} = 1$ *implies* $\Delta\bar{A} = 1$.

II. *The components* $\underline{A}^i$ *and* A^i, $i = 1, \ldots, r$ *of* $\underline{\mathbb{A}}$ *and* $\mathbb{A}$ *are related as* $\underline{A}^i_t = \int_0^t \lambda^i dA^i$ *where* $\operatorname{col}\{\lambda^1, \ldots, \lambda^r\} = \Lambda = \{\Lambda_t, \mathcal{F}_t\}$ *is a nonnegative predictable process such that the associated Hellinger process is bounded:* $\mathcal{H}_t = \int_0^t \sum_{i=1}^r (\sqrt{d\underline{A}^i} - \sqrt{dA^i})^2 + \sum_{\substack{s \leqslant t \\ 0<\Delta A_s<1}} (\sqrt{1-\Delta\underline{\bar{A}}_s} - \sqrt{1-\Delta\bar{A}_s})^2 < \infty$.

2) *Assume* $P \ll P$, *and denote by* z_t *a right-continuous modification of the martingale* $E(d\underline{P}/dP|\mathcal{F}_t)$ $0 \leqslant t \leqslant 1$. *Then* $z_t = \exp\{m_t + \sum_{s \leqslant t} \Phi_1(\Delta m_s)\}$ *where* $\Phi_1(x) = \ln(1+x) - x$, *and*

$$m_t = \int_0^t (\Lambda - \mathbb{I}_r)^T d\mathfrak{M} \tag{3.2}$$

REMARK 3.3. The process $z = (z_t, \mathcal{F}_t, P)$, being a nonnegative supermartingale with $E(z|P) = 1$ as well as a local martingale, is a solution of the Doleans-Dade equation $z_t = 1 + \int_0^t z_{s-} dm_s$, $0 \leqslant t \leqslant 1$ (LIPTSER AND SHIRYAYEV [18], p. 288, or GILL AND JOHANSEN [8]), hence $z_t = \mathcal{E}(m)_t$.

REMARK 3.4. By (3.1) and (3.2) we have

$$\Delta m = (\Lambda - \mathbb{I})^T \Delta \mathbb{N} + (1 - \Delta \bar{N}) \left[(1 - \Delta \underline{\bar{A}})(1 - \Delta \bar{A})^{-1} - 1 \right] \tag{3.3}$$

and

$$\Phi_1(\Delta m) = \Phi_1^T (\Lambda - \mathbb{I}_r) \Delta \mathbb{N} + (1 - \Delta \bar{N}) \Phi_1 \left[(1 - \Delta \underline{\bar{A}})(1 - \Delta \bar{A})^{-1} - 1 \right] \tag{3.4}$$

with $\Phi_1(x) = \mathrm{col}\{\Phi_1(x^i),\ i = 1, \ldots, r\}$ for $x = \mathrm{col}\{x^1, \ldots, x^r\}$. Hence $z_t = \exp\{\int_0^t \ln^T \Lambda d\mathbb{N} - \underline{\bar{A}}_t^c + \bar{A}_t^c + \sum_{s \leqslant t} (1 - \Delta \bar{N}_s) \ln (1 - \Delta \underline{\bar{A}}_s)(1 - \Delta \bar{A}_s)^{-1}\}$ (cf. LIPTSER AND SHIRYAYEV [18], p. 312).

REMARK 3.5. By (3.3)

$$\eta_t \equiv \sum_{s \leqslant t} (1 - \sqrt{1 + \Delta m_s})^2 \tag{3.5}$$

$$= \int_0^t \mathbb{U}^T \mathrm{diag} d\mathbb{N} \mathbb{U} + \sum_{s \leqslant t} (1 - \Delta \bar{N}_s) \left[(1 - \Delta \underline{\bar{A}}_s)^{\frac{1}{2}} (1 - \Delta \bar{A}_s)^{-\frac{1}{2}} - 1 \right]^2$$

with $\mathbb{U} = \mathrm{col}\{\sqrt{\lambda^i} - 1 = \sqrt{dA^i / dA^i} - 1, i = 1, \ldots, r\}$. The compensator of this process coincides with the Hellinger process, $\tilde{\eta} = \bar{\mathcal{H}}$.

REMARK 3.6. It is interesting to note that the class of 'alternative' compensators $\underline{\mathbb{A}}$ is restricted to those for which $\underline{\mathbb{A}} - \mathbb{A}$ is dominated by $<\mathbb{M}>$ in the sense that for a certain r-vector valued predictable process $\mathbb{H}$

$$\underline{\mathbb{A}}_t - \mathbb{A}_t = \int_0^t \mathbb{H}^T d<\mathbb{M}> .$$

If $\underline{P} \ll P$ then $z = \mathcal{E}(m)$ with $m_t = \int_0^t \mathbb{H}^T d\mathbb{M}$. Obviously, $\mathbb{H} = (V^{-1})^T (\Lambda - \mathbb{I})$ and $\underline{\mathbb{A}} - \mathbb{A} = <\mathbb{M}, m>$.

3.3. Here we give a useful representation for the likelihood ratio process, to be used in the next section.

LEMMA 3.3. *Let $\underline{P} \ll P$. Then*

$$z = \exp\{2\mathfrak{m}(\mathbb{U}) - 2\mathfrak{K} + R\} \tag{3.6}$$

where

$$\mathfrak{m}(\mathbb{U})_t = \int_0^t \mathbb{U}^T d\mathfrak{N} \tag{3.7}$$

is a local square integrable martingale with

$$<\mathfrak{m}(\mathbb{U})>_t = \int_0^t \mathbb{U}^T \mathrm{diag} d<\mathfrak{N}> \mathbb{U} < \infty \quad P \text{ a.s.}, \tag{3.8}$$

while

$$R_t = 2\sum_{s\leqslant t}\Phi_2(\sqrt{1+\Delta m_s}-1) + 2[\overline{\mathfrak{N}},\mathfrak{K}]_t - \int_0^t \mathbb{U}^T \mathrm{diag} d\mathfrak{N}\mathbb{U} \tag{3.9}$$

with

$$\Phi_2(x) = \ln(1+x)-x+\tfrac{1}{2}x^2 .$$

PROOF. By (3.2)

$$m_t = 2\int_0^t \mathbb{U}^T d\mathfrak{N} + \int_0^t \mathbb{U}^T \mathrm{diag} d\mathfrak{N}\mathbb{U} . \tag{3.10}$$

By (3.4) and (3.5)

$$\tfrac{1}{2}\sum_{s\leqslant t}\Phi_1(\Delta m_s) = \sum_{s\leqslant t}\Phi_2(\sqrt{1+\Delta m_s}-1) - \mathfrak{K}_t - (\eta_t - \tilde{\eta}_t) , \tag{3.11}$$

since $\frac{1}{2}\Phi_1(x-1) = \Phi_2(\sqrt{x}-1)-(\sqrt{x}-1)^2$ and $\mathfrak{K} = \tilde{\eta}$; obviously,

$$\eta_t - \tilde{\eta}_t = \int_0^t \mathbb{U}^T \mathrm{diag} d\mathbb{M}\mathbb{U} + \sum_{s\leqslant t}\Delta\overline{\mathfrak{N}}(\sqrt{1-\Delta\overline{A}} - \sqrt{1-\Delta\overline{A}})^2 . \tag{3.12}$$

Now (3.6) easily follows from (3.9) - (3.12), taking into account that $\mathfrak{N} = \mathbb{M} + [\mathbb{A},\overline{\mathfrak{N}}]$ by definition.

By Assertion 2) of Lemma 3.2

$$\begin{aligned} <\mathfrak{m}(\mathbb{U})>_t &= \int_0^t \mathbb{U}^T \mathrm{diag} d\mathbb{A}\mathbb{U} + \sum_{s\leqslant t}\frac{(\mathbb{U}_s^T\Delta\mathbb{A}_s)^2}{1-\Delta\overline{A}_s} \\ &\leqslant \int_0^t \mathbb{U}^T \mathrm{diag} d\mathbb{A}\mathbb{U} + \sum_{s\leqslant t}\frac{\Delta\overline{A}_s}{1-\Delta\overline{A}_s}\mathbb{U}_s^T \mathrm{diag}\Delta\mathbb{A}_s\mathbb{U}_s \\ &\leqslant \int_0^t \mathbb{U}^T \mathrm{diag} d\mathbb{A}\mathbb{U} + \sum_{s\leqslant t}\mathbb{U}_s^T \mathrm{diag}\Delta\mathfrak{C}_s\mathbb{U}_s < \infty \ P \ a.s. \end{aligned} \tag{3.13}$$

Here we first used the Schwartz inequality and then the boundedness of the Hellinger process.[1] Hence (3.8) holds.

□

4. LAN FOR COUNTING PROCESSES

4.1. Let $\{\Omega^n, \mathcal{F}^n, (\mathcal{F}_t^n, 0\leqslant t\leqslant 1), P^n)\}$, $n = 1,2,...$ be a sequence of stochastic bases of the same type as above. Let $\mathbb{N}^n = (\mathbb{N}_t^n, \mathcal{F}_t^n, P^n)$ be an r_n-variate counting process with the Doob-Meyer decomposition $\mathbb{N}^n = \mathbb{M}^n + \mathbb{A}^n$, where r_n, $n = 1,2,...$ is a nondecreasing sequence of integers.

Define also $\mathfrak{N}_t^n = \int_0^t (V^n)^{-1} d\mathbb{M}^n$ where $V^n = I_{r_n} - \Delta\mathbb{A}^n \otimes \mathbb{I}_{r_n}$.

Let $\mathbb{H}^n = \{\mathbb{H}_t^n, \mathcal{F}_t^n, P^n\}$, $n = 1,2,...$ be a sequence of r_n-verctor valued predictable processes such

1. Use also the following inequalities : $\sum_{s\leqslant t} I(0<\Delta\overline{A}\leqslant\frac{1}{2})\ \mathbb{U}_s^T \mathrm{diag}\Delta\mathfrak{C}_s\mathbb{U}_s \leqslant 2\sum_{s\leqslant t}\mathbb{U}_s^T\mathrm{diag}\Delta\mathbb{A}_s\mathbb{U}_s$ and $\sum_{s\leqslant t} I(\frac{1}{2}\leqslant\Delta\overline{A}<1)\ \mathbb{U}_s^T\mathrm{diag}\Delta\mathfrak{C}_s\mathbb{U}_s \leqslant C\sum_{s\leqslant t}\mathbb{U}_s^T\mathrm{diag}\Delta\mathbb{A}_s\mathbb{U}_s$ with a certain constant C determined by the fact that the number of jumps of $\overline{A}_s$, $s\leqslant t$, exceeding $\frac{1}{2}$ is finite.

that

$$\mathrm{m}(\mathbb{H}^n)_t = \int_0^t \mathbb{H}^{nT} d\mathfrak{M}^n \ , \ n = 1,2,... \tag{4.1}$$

is a sequence of local square integrable martingales.

By Corollary 2 of LIPTSER AND SHIRYAYEV [19] this sequence is asymptotically normal (see Theorem 4.1 below) under the following Conditions $\mathbb{H}$:

$\mathbb{H}.1$. For each t, $0 \leqslant t \leqslant 1$ and ϵ, $0 < \epsilon \leqslant 1$

$$\int_0^t \mathbb{H}^{nT}_{(>\epsilon)} \mathrm{diag} d\mathbb{A}^n \mathbb{H}^n_{(>\epsilon)} + \sum_{s \leqslant t} I(|\mathbb{H}^{nT}\Delta\mathfrak{C}^n| > \epsilon)(1 - \Delta\bar{A}^n)(\mathbb{H}^{nT}\Delta\mathfrak{C}^n)^2 \to 0 \tag{4.2}$$

in P^n probability as $n \to \infty$, where

$$\mathbb{H}^n_{(>\epsilon)} = \mathrm{col}\{I(|H^{in}| > \epsilon)H^{in} \ , \ i = 1, \ldots, r_n\} \ , \ \mathbb{H}^n = \mathrm{col}\{H^{in} \ , \ i = 1, \ldots, r_n\} \tag{4.3}$$

$\mathbb{H}.2$. For each t, $0 \leqslant t \leqslant 1$

$$<\mathrm{m}(\mathbb{H}^n)>_t = \int_0^t \mathbb{H}^{nT} \mathrm{diag} d\mathbb{A}^n \mathbb{H}^n + \sum_{s \leqslant t} (1 - \Delta\bar{A}^n)(\mathbb{H}^{nT}\Delta\mathfrak{C}^n)^2 \to <W> \tag{4.4}$$

where $W = (W_t, \mathfrak{F}_t)_{0 \leqslant t \leqslant 1}$ is a continuous Gaussian martingale with quadratic variation $<W> = [W] = EW^2$, a nondecreasing continuous deterministic function (cf. GREENWOOD AND SHIRYAYEV [9], § 5.2).

THEOREM 4.1. *Under the Conditions* $\mathbb{H}$

$$\mathrm{m}(\mathbb{H}^n) \xrightarrow{d(P^n)} W \tag{4.5}$$

in the sense of the weak convergence in $\mathfrak{D}([0,1])$ *with resp. to* P^n *(cf.* GREENWOOD AND SHIRYAYEV [9], § 2.2).

REMARK 4.1. For checking the above statement take into consideration that the integer valued random measure μ^n, associated to $\mathrm{m}(\mathbb{H}^n)$ by

$$\mu^n((0,t], \Gamma) = \sum_{s \leqslant t} I(\Delta\mathrm{m}(\mathbb{H}^n)_s \in \Gamma), \ \Gamma \in \mathfrak{B}(R_0), \ R_0 = R \setminus \{0\}$$

with

$$\Delta\mathrm{m}(\mathbb{H}^n) = \mathbb{H}^{nT}\Delta\mathbb{N}^n - (1 - \Delta\bar{N})\mathbb{H}^{nT}\Delta\mathfrak{C}^n \ ,$$

is such that

$$\int_0^t \int_{|x|>\epsilon} x^2 \mu^n(ds, dx) = \sum_{s \leqslant t} \Delta\mathrm{m}(\mathbb{H}^n)_s^2 I(|\Delta\mathrm{m}(\mathbb{H}^n)_s| > \epsilon) \tag{4.6}$$

$$= \int_0^t \mathbb{H}^{nT}_{(>\epsilon)} \mathrm{diag} d\mathbb{N}^n \mathbb{H}^n_{(>\epsilon)} + \sum_{s \leqslant t} (1 - \Delta\bar{N}^n)(\mathbb{H}^{nT}\Delta\mathfrak{C}^n)^2 I(|\mathbb{H}^{nT}\Delta\mathfrak{C}^n| > \epsilon)$$

Here we have used the following simple relation :

$$I(|\Delta\mathrm{m}(\mathbb{H}^n)| > \epsilon)\Delta\mathrm{m}(\mathbb{H}^n) = \mathbb{H}^{nT}_{(>\epsilon)}\Delta\mathbb{N}^n - (1 - \Delta\bar{N}^n) I(|\mathbb{H}^{nT}\Delta\mathfrak{C}^n| > \epsilon)\mathbb{H}^{nT}\Delta\mathfrak{C}^n \ . \tag{4.7}$$

Now, we can easily see that on the left hand side of (4.2) we have the compensator of the expression (4.6). Hence, denoting the compensator of μ^n by ν^n, one can rewrite (4.2) as follows :

$$\int_0^t \int_{|x|>\epsilon} x^2 d\nu^n(ds, dx) \to 0 \ .$$

Below we will need the following simple corollary of theorem 4.1.

COROLLARY 4.1. *Let a sequence* $\mathbb{H}^n$, $n = 1,2,\ldots$ *of* r_n*-valued predictable processes satisfy the following Conditions* $\mathbb{H}'$ *: for each* t, $0 \leqslant t \leqslant 1$

$\mathbb{H}'.0.$ $\quad \int_0^t \mathbb{H}^{nT} d\mathbb{A}^n \to 0$;

$\mathbb{H}'.1.$ $\quad \int_0^t \mathbb{H}^{nT}_{(>\epsilon)} \mathrm{diag} d\mathbb{A}^n \mathbb{H}^n_{(>\epsilon)} \to 0$, $0 < \epsilon \leqslant 1$;

$\mathbb{H}'.2.$ $\quad \int_0^t \mathbb{H}^{nT} \mathrm{diag} d\mathbb{A}^n \mathbb{H}^n \to \langle W \rangle_t$

in P^n *probability as* $n \to \infty$. *Then*

$$\int_0^t \mathbb{H}^{nT} d\mathbb{N}^n \xrightarrow{d(P^n)} W_t \, .$$

4.2. Suppose that a probability measure $\underline{P}^n$ in addition to P^n is given on a measurable space $\{\Omega^n, \mathcal{F}^n\}$ of the preceding subsection. Suppose in addition that the filtration $\{\mathcal{F}_t^n, 0 \leqslant t \leqslant 1\}$ is minimal: $\mathcal{F}_t^n = \sigma\{\mathbb{N}_s^n : s \leqslant t\}$ where $\mathbb{N}^n = (\mathbb{N}_t^n, \mathcal{F}_t^n, P^n)$ is an r_n-variate counting process with the compensator $\mathbb{A}^n = (\mathbb{A}_t^n, \mathcal{F}_t^n, P^n)$. Let $\underline{\mathbb{N}}^n = (\mathbb{N}_t^n, \mathcal{F}_t^n, \underline{P}^n)$ be another counting process with the compensator $\underline{\mathbb{A}}^n = (\underline{\mathbb{A}}_t^n, \mathcal{F}_t^n, \underline{P}^n)$.

For each n assume $\underline{P}^n \ll P^n$ and, in accordance with II of Theorem 3.1, define the Hellinger process

$$\mathcal{H}_t^n = \int_0^t \mathbb{U}^{nT} \mathrm{diag} d\mathbb{A}^n \mathbb{U}^n + \sum_{\substack{s \leqslant t \\ 0 < \Delta A_s^n < 1}} (\sqrt{1 - \Delta \underline{A}_s^n} - \sqrt{1 - \Delta A_s^n})^2 \tag{4.8}$$

where

$$\mathbb{U}^n = \mathrm{col}\{U^{in} = \sqrt{d\underline{A}^{in}/dA^{in}} - 1, i = 1, \ldots, r_n\} \, .$$

Obviously,

$$\Lambda^n = \mathrm{col}\{\lambda^{in} = (U^{in} + 1)^2, i = 1, \ldots, r_n\} \, .$$

Let the following Conditions $\mathcal{H}$ be satisfied :

$\mathcal{H}1$. For each t, $0 \leqslant t \leqslant 1$ and ϵ, $0 < \epsilon \leqslant 1$

$$\tilde{\eta}^n_{(>\epsilon)t} \equiv \int_0^t \hat{\mathbb{U}}^{nT}_{(>\epsilon)} \mathrm{diag} d\mathbb{A}^n \hat{\mathbb{U}}^n_{(>\epsilon)} \tag{4.9}$$

$$+ \sum_{s \leqslant t} I(|\Delta \underline{A}_s^n - \Delta A_s^n| > \epsilon(1 - \Delta A_s^n))(\sqrt{1 - \Delta \underline{A}_s^n} - \sqrt{1 - \Delta A_s^n})^2 \to 0$$

in P^n probability as $n \to \infty$, where

$$\hat{\mathbb{U}}^n_{(>\epsilon)} = \mathrm{col}\{I(|\lambda^{in} - 1| > \epsilon) U^{in}, \; i = 1, \ldots, r_n\} \tag{4.10}$$

$\mathcal{H}2$. For each t, $0 \leqslant t \leqslant 1$

$$\mathcal{H}^n \to \frac{1}{4} \langle W \rangle \tag{4.11}$$

in P^n probability as $n \to \infty$.

PROPOSITION 4.1.
(i) Under the Conditions $\mathcal{K}$

$$\mathrm{m}(\mathbb{U}^n) \xrightarrow{d(P^n)} \frac{1}{4}W \tag{4.12}$$

(cf. (3.7), (4.1) and (4.5) with $\mathbb{H} = 2\mathbb{U}$ *)*
(ii) The Conditions $\mathcal{K}$ *are equivalent to the Conditions* $\mathbb{U}$ *defined by (4.2) - (4.4) for the special case of* $\mathbb{H} = 2\mathbb{U}$.

REMARK 4.2. As the Conditions $\mathbb{U}$ are those of Theorem 4.1 for the special case of $\mathbb{H} = 2\mathbb{U}$, the assertion of Lemma 3.3 concerning the process (3.7) allows us to deduce Assertion (i) of Proposition 4.1 directly from Assertion (ii) and Theorem 4.1. The Assertion (ii) will be proved below.

REMARK 4.3. Notice the difference between $\hat{\mathbb{U}}^n_{(>\epsilon)}$ given by (4.10) and

$$\mathbb{U}^n_{(>\epsilon)} = \mathrm{col}\{I(|U^{in}|>\epsilon)U^{in}\ ,\ i = 1, \ldots, r_n\} \tag{4.13}$$

(cf. (4.3)). However using the simple inequalities

$$I(|\sqrt{1+x}-1|>\epsilon)\leqslant I(|x|>\epsilon)$$

and

$$I(|x-y|>\epsilon)\leqslant I(|x|>\epsilon/2) + I(|y|>\epsilon/2)\,.$$

we get

$$\int_0^t \mathbb{U}^{nT}_{(>\epsilon)}\,\mathrm{diag}d\mathbb{A}^n\mathbb{U}^n_{(>\epsilon)} \leqslant \int_0^t \hat{\mathbb{U}}^{nT}_{(>\epsilon)}\mathrm{diag}d\mathbb{A}^n\hat{\mathbb{U}}^n_{(>\epsilon)} \tag{4.14}$$

$$\leqslant \int_0^t \mathbb{U}^{nT}_{(>\epsilon/4)}\mathrm{diag}d\mathbb{A}^n\mathbb{U}^n_{(>\epsilon/4)} + \int_0^t \mathbb{U}^{nT}_{(>\sqrt{\epsilon/2})}\mathrm{diag}d\mathbb{A}^n\mathbb{U}^n_{(>\sqrt{\epsilon/2})}\,.$$

Proof of Assertion (ii) of Proposition 4.1. We proceed in three steps. In step 1) we show that the Conditions $\mathcal{K}$ imply (4.2) with $\mathbb{H} = \mathbb{U}$. In step 2) we show that the Conditions $\mathbb{U}$ imply Condition $\mathcal{K}1$. In conclusion, it is shown in step 3) that the difference between $\mathcal{K}^n$ and the left hand side of (4.4) with $\mathbb{H} = 2\mathbb{U}$ vanishes as $n\to\infty$ under the Conditions $\mathcal{K}$, as well as under the Conditions $\mathbb{U}$.

1) By (4.14), under the Conditions $\mathcal{K}$ the first term on the left hand side of (4.2) with $\mathbb{H} = \mathbb{U}$ tends to zero in P^n probability as $n\to\infty$. We will show that the same holds for the second term, as well. The latter term does not exceed

$$\sum_{s\leqslant t} I(|(\Lambda^n_s - \mathbb{I}_{r_n})^T\Delta\mathcal{Q}^n_s|>\epsilon)(1-\Delta\bar{A}^n_s)\{(\Lambda^n_s - \mathbb{I}_{r_n})^T\Delta\mathcal{Q}^n_s\}^2 \tag{4.15}$$

$$+ \sum_{s\leqslant t}(1-\Delta\bar{A}^n_s)(\mathbb{U}^{nT}_s\mathrm{diag}\Delta\mathcal{Q}^n_s\mathbb{U}^n_s)^2\ ,$$

as is easily seen by applying the simple inequality

$$|x-y|I(|x-y|>\epsilon)\leqslant 4|x|^2I(|x|>\epsilon/2) + 4|y|^2I(|y|>\epsilon/2) \tag{4.16}$$

(see ANDERSEN AND GILL [1], p. 1107) to

$$(\Lambda^n - \mathbb{I}_{r_n})\Delta\mathcal{Q}^n = 2\mathbb{U}^{nT}\Delta\mathcal{Q}^n + \mathbb{U}^{nT}\mathrm{diag}\Delta\mathcal{Q}^n\mathbb{U}^n \tag{4.17}$$

$$= 1-(1-\Delta\bar{A}^n)/(1-\Delta\bar{A}^n)\,.$$

Since for each $\epsilon>0$ one can choose a constant C that ensures the inequality $|x|\leqslant C(\sqrt{1+x}-1)^2$

whenever $|x|>\epsilon$ (e.g. via $x^2/(1+|x|) \asymp (\sqrt{1+x}-1)^2$; KABANOV et al. [14], p. 644), the expression (4.15) in turn does not exceed

$$C\{\sum_{s\leqslant t} I(|(\Lambda_s^n - \mathbb{I}_{r_n})^T \Delta\mathcal{C}_s^n|>\epsilon)(1-\Delta\bar{A}_s^n)(\sqrt{1-(\Lambda_s^n-\mathbb{I}_{r_n})^T\Delta\mathcal{C}_s^n}-1)^2\}^2 \tag{4.18}$$
$$+ \sup_{s\leqslant t}\mathbb{U}_s^{nT}\mathrm{diag}\Delta\mathbb{A}_s^n\mathbb{U}_s^n\cdot\sum_{s\leqslant t}\mathbb{U}_s^{nT}\mathrm{diag}\Delta\mathcal{C}_s^n\mathbb{U}_s^n .$$

By (4.17) and Condition $\mathcal{K}1$, the first term in (4.18) tends to zero in P^n probability as $n\to\infty$. In view of the last inequality in (3.13) and the fact that[1]

$$\sup_{s\leqslant t}\mathbb{U}_s^{nT}\mathrm{diag}\Delta\mathbb{A}_s^n\mathbb{U}_s^n \leqslant \sup_{s\leqslant t}\Delta\mathcal{K}_s^n \to 0 \text{ in } P^n \text{ probability as } n\to\infty, \tag{4.19}$$

the second term in (4.18) vanishes as well. Thus (4.2) for $\mathbb{H} = \mathbb{U}$ is proved.

2) Let the Conditions $\mathbb{U}$ hold ; By (4.14) again, it suffices to bound the second term of $\tilde{\eta}^n$ (see (4.9)) and to show that it vanishes as $n\to\infty$. By the simple inequality $|\sqrt{1+x}-1|\leqslant|x|$, this term does not exceed

$$\sum_{s\leqslant t} I\left(|(\Lambda_s^n-\mathbb{I}_{r_n})^T\Delta\mathcal{C}_s^n|>\epsilon\right)(1-\Delta\bar{A}_s^n)\left\{(\Lambda_s^n-\mathbb{I}_{r_n})^T\Delta\mathcal{C}_s^n\right\}^2$$
$$\leqslant 4\sum_{s\leqslant t} I\left(|\mathbb{U}_s^{nT}\Delta\mathcal{C}_s^n|>\epsilon/4\right)\{2\mathbb{U}_s^{nT}\Delta\mathcal{C}_s^n\}^2(1-\Delta\bar{A}_s^n)$$
$$+4\sup_{s\leqslant t}\mathbb{U}_s^{nT}\mathrm{diag}\Delta\mathbb{A}_s^n\mathbb{U}_s^n\cdot\sum_{s\leqslant t}\mathbb{U}_s^{nT}\mathrm{diag}\Delta\mathcal{C}_s^n\mathbb{U}_s^n \ ; \tag{4.20}$$

here we have used (4.16) and (4.17). The second term on the right hand side of (4.20) tends to zero by the same arguments as above (cf. the similar term in (4.18)) ; so does the first term as well, by (4.2) for $\mathbb{H} = \mathbb{U}$. Thus (4.9) is proved.

3) In view of the assertions proved in steps 1) and 2), all we need is that

$$\sum_{s\leqslant t} I\left(|(\Lambda_s^n-\mathbb{I}_{r_n})^T\ \Delta\mathcal{C}_s^n|\leqslant\epsilon\right)(1-\Delta\bar{A}_s^n)|\sqrt{1-(\Lambda_s^n-\mathbb{I}_{r_n})^T\Delta\mathcal{C}_s^n}-1)^2 - \tag{4.21}$$
$$\left\{\frac{1}{2}(\Lambda_s^n-\mathbb{I}_{r_n})^T\Delta\mathcal{C}_s^n\right\}^2|\to 0$$

in P^n probability as $n\to\infty$, either under the Conditions $\mathcal{K}$ or $\mathbb{U}$.

Since $1-\sqrt{1-x} = x/2+x^2/8+\mathcal{O}(x^3)$ for sufficiently small values of x, a constant C can be chosen such that

$$|(1-\sqrt{1-x})^2-(\tfrac{1}{2}x)^2|\leqslant C|x^3|$$

Applying this inequality to the left-hand side of (4.21), one can see that it does not exceed

$$C\epsilon\sum_{s\leqslant t}(1-\Delta\bar{A}_s^n)\left\{(\Lambda_s^n-\mathbb{I}_{r_n})^T\Delta\mathcal{C}_s^n\right\}^2 \leqslant$$
$$2C\epsilon\sum_{s\leqslant t}(1-\Delta\bar{A}_s^n)(2\mathbb{U}_s^{nT}\Delta\mathcal{C}_s^n)^2$$

1. As it has been mentioned in Remark 2.1, by Lemma 1 of MCLEISH [20], p. 146 from Condition *S.II* and the continuity of $\mathcal{K}^W$ follows $\sup_{s\leqslant t}|\mathcal{K}_s^n-\mathcal{K}_s^W|\to 0$ and, in particular $\sup_{s\leqslant t}|\Delta\mathcal{K}_s^n|\to 0$ in P^n probability as $n\to\infty$; cf. GREENWOOD AND SHIRYAYEV [9], p. 105.

$$+2C\epsilon\sup_{s\leqslant t}\mathbb{U}_s^{nT}\mathrm{diag}\Delta\mathbb{A}_s^n\mathbb{U}_s^n\cdot\sum_{s\leqslant t}\mathbb{U}_s^{nT}\mathrm{diag}\Delta\mathcal{C}_s^n\mathbb{U}_s^n\ .$$

This and (3.13) imply (4.21). The concluding step 3) is proved.

□

4.3. The next three lemmas establish asymptotic negligability of the remainder term R (see (3.9)) in the representation (3.6).

LEMMA 4.1. *Under the Conditions* $\mathfrak{K}$, *for each* t, $0\leqslant t\leqslant 1$

$$\sup_{s\leqslant t}R_s^{(1)}\to 0\ ,\ R_t^{(1)}=\int_0^t\mathbb{U}^{nT}\mathrm{diag}d\mathfrak{M}^n\mathbb{U}^n\ , \tag{4.22}$$

in P^n *probability as* $n\to\infty$.

PROOF . By Assertion (ii) of Proposition 4.1 and (4.2) with $\mathbb{H}=\mathbb{U}$, the arguments indicated in the footnote on p. 42 lead to

$$\int_0^t\mathbb{U}_{(>\epsilon)}^{nT}\mathrm{diag}d\mathcal{C}^n\mathbb{U}_{(>\epsilon)}^n\to 0 \text{ in } P^n \text{ probability as } n\to\infty.$$

Thus, it suffices to prove (4.22) with $\mathbb{U}_{(\leqslant\epsilon)}^n=\mathbb{U}^n-\mathbb{U}_{(>\epsilon)}^n$ in place of $\mathbb{U}^n$. But this is a direct consequence of Assertion (i) of Proposition 4.1, as

$$\int_0^t\mathbb{U}_{(\leqslant\epsilon)}^{nT}\mathrm{diag}d\mathfrak{M}^n\mathbb{U}_{(\leqslant\epsilon)}^n\leqslant\epsilon\ \mathfrak{m}(\mathbb{U}^n)_t$$

□

LEMMA 4.2. *Under the Conditions* $\mathfrak{K}$, *for each* t, $0\leqslant t\leqslant 1$

$$\sup_{s\leqslant t}|R_s^{(2)}|\to 0\ ,\ R^{(2)}=[\overline{\mathfrak{M}}^n,\ \mathfrak{K}^n] \tag{4.23}$$

in P^n *probability as* $n\to\infty$.

PROOF . By Assertion 2) of Lemma 3.2, (4.19) and the boundedness of the Hellinger process

$$\begin{aligned}<R^{(2)}>_t&=\sum_{s\leqslant t}(\Delta\mathfrak{K}_s^n)^2\frac{\Delta\overline{A}_s^n}{1-\Delta\overline{A}_s^n} \qquad (4.24)\\ &\leqslant\sup_{s\leqslant t}\Delta\mathfrak{K}_s^n\cdot\sum_{s\leqslant t}\frac{\Delta\mathfrak{K}_s^n}{1-\Delta\overline{A}_s^n}\\ &\leqslant\sup_{s\leqslant t}\Delta\mathfrak{K}_s^n\cdot C\mathfrak{K}_t^n\to 0\end{aligned}$$

in P^n probability as $n\to\infty$ (a constant C is defined by the arguments indicated in the footnote on p. 42). Obviously, (4.23) is implied by (4.24).

□

LEMMA 4.3. *Under the Conditions* $\mathfrak{K}$, *for each* t, $0\leqslant t\leqslant 1$

$$\sup_{s\leqslant t}|R_s^{(3)}|\to 0,\ R_t^{(3)}=\sum_{s\leqslant t}\Phi_2(\sqrt{1+\Delta m_s^n}-1) \tag{4.25}$$

in P^n *probability as* $n\to\infty$.

REMARK 4.3. The last assertion is the special case of Assertion 1.B for $i=3$, in DZHAPARIDZE [6], Subsection 1.2. In fact

$$R_t^{(3)} = \int_0^t \Phi_2^T(\mathbb{U}^n)d\mathbb{N}^n + \sum_{s\leqslant t}(1-\Delta\overline{N}_s^n)\Phi_2((1-\Delta\underline{\overline{A}}^n)^{\frac{1}{2}}(1-\overline{A}^n)^{-\frac{1}{2}}-1)$$

$$= \int_0^t\int_{R_0}\Phi_2(\sqrt{1+x}-1)d\mu^n$$

with μ^n defined as in Remark 4.1 for the particular choice of $\mathbb{H}^n$, namely, $\mathbb{H}^n = \Lambda^n - \mathbb{I}_{r_n}$ (cf. DZHAPARIDZE [6], p. 16).

4.4. In conclusion, let us formulate the principal results of this section - Theorem 4.2 and its Corollary 4.2 stating the LAN for counting processes.

THEOREM 4.2. *Under the Conditions $\mathcal{K}$ the following two statements hold:*

(i) $z^n \xrightarrow{d(P^n)} \exp\{W-\frac{1}{2}<W>\}$,

(ii) $z^n \xrightarrow{d(P^n)} \exp\{W+\frac{1}{2}<W>\}$,

PROOF. Assertion (ii) is derived from Assertion (i) by the arguments used in GREENWOOD AND SHIRYAYEV [9] (see the proof of Statement 3 of Theorem 8 on p. 99).

As for Assertion (i), it follows directly from Lemma 3.3, Proposition 4.1 and the Lemmas 4.1 - 4.3. □

COROLLARY 4.2. *Let the Conditions $\mathcal{K}$ are satisfied. Then the following two statements hold :*
(i) For each t, $0\leqslant t\leqslant 1$

$$z^n = \exp\{\mathrm{m}(2\mathbb{U}^n)-\frac{1}{2}<W>+r^n\}$$

where a remainder term r^n is such that

$$\sup_{s\leqslant t}|r_s^n|\to 0 \tag{4.26}$$

both in P^n and $\underline{P}^n$ probability as $n\to\infty$, while the first term $\mathrm{m}(2\mathbb{U}^n)$ is asymptotically normal :

$$\mathrm{m}(2\mathbb{U}^n) \xrightarrow{d(P^n)} W \tag{4.27}$$

and

$$\mathrm{m}(2\mathbb{U}^n) \xrightarrow{d(P^n)} W+<W> \tag{4.28}$$

(ii) Let $\mathbb{S}^n = \{\mathbb{S}_t^n, \mathcal{G}_t^n, P^n\}$, $n = 1, 2,...$ be a sequence of r_n-variate predictable processes such that for some unboundedly increasing sequence of numbers k_n, $n = 1, 2,...$ it satisfies the Conditions $\mathbb{H}$ with $\mathbb{H}^n = k_n^{-\frac{1}{2}}\mathbb{S}^n$. Besides, for each t, $0\leqslant t\leqslant 1$

$$\int_0^t(\mathbb{U}^n-k_n^{-\frac{1}{2}}\mathbb{S}^n)^T\mathrm{diag}d\mathbb{A}^n(\mathbb{U}^n-k_n^{-\frac{1}{2}}\mathbb{S}^n)\to 0 \tag{4.29}$$

in P^n probability as $n\to\infty$. Then

$$z^n = \exp\{2k_n^{-\frac{1}{2}}\mathrm{m}(\mathbb{S}^n)-\frac{1}{2}<W>+r'^n\} \tag{4.30}$$

where a remainder term r'^n and the first term $2k_n^{-\frac{1}{2}}\mathrm{m}(\mathbb{S}^n)$ satisfy (4.26) and, respectively (4.27) and

(4.28) with $k_n^{-\frac{1}{2}}\mathbb{S}^n$ *in place of* $\mathbb{U}^n$.

Finally, if $k_n^{-\frac{1}{2}}\mathbb{S}^n$ *satisfies the Conditions* $\mathbb{H}'$ *then the first term in (4.30) is simplified to* $2k_n^{-\frac{1}{2}}\int_0^t \mathbb{S}^{nT}d\mathbb{N}^n$.

PROOF. Obviously, for the proof of Assertion (i) it suffices to check (4.26) for

$$r^n = R^n - 2(\mathfrak{K}^n - \tfrac{1}{4}<W>)$$

(see (3.6) and (3.9)), in view of Lemmas 4.1 - 4.3 and the footnote on p. 46.

As for Assertion (ii), we apply Theorem 4.1 and its Corollary 4.1, and then the fact that (4.29) implies

$$<\mathfrak{m}(\mathbb{U}^n - k_n^{-\frac{1}{2}}\mathbb{S}^n)> \to 0$$

in P^n probability as $n\to\infty$, each t, $0\leqslant t\leqslant 1$, which is established by using the arguments leading to (3.13). The latter fact ensures the property (4.26) for

$$r'^n = r^n + 2\mathfrak{m}(\mathbb{U}^n - k_n^{-\frac{1}{2}}\mathbb{S}^n)\,.$$

□

5. CONSISTENCY AND ASYMPTOTIC NORMALITY.

5.1. Consider the situation described in Subsection 2.1, in which for each $n = 1,2,\ldots$ the r-variate counting process $\mathbb{N}^n$ possesses the pair $(\bar{A}_t^n, \Psi_t^n)_\beta$ of $(P_\beta^n, \mathcal{F}_t^n)$-predictable characteristics satisfying Condition I. Suppose that the r_n-vector-valued process $\sqrt{\Psi^n(\beta)}$ is continuously differentiable (in β) in the following sense:

CONDITION II. There is a sequence of continuous in β r_n-vector valued predictable processes

$$\mathbb{L}^n(\beta)\ (\equiv \frac{\partial}{\partial\beta}\sqrt{\Psi^n(\beta)}) = \{\mathbb{L}_t^n(\beta), \mathcal{F}_t^n, P^n\}\,,\ n = 1,2,\ldots$$

such that

II 1. For each real valued b such that $\underline{\beta}^n = \beta + b/\sqrt{k_n} \in \mathcal{B}$, eventually, with the same sequence of numbers $k_n, n = 1,2,\ldots$ as in Condition I,

$$\int_0^1 |\sqrt{\Psi^n(\underline{\beta}^n)} - \sqrt{\Psi^n(\beta)} - \frac{b}{\sqrt{k_n}}\mathbb{L}^n(\beta)|^2 d\bar{A}^n \to 0 \quad \text{in } P_\beta^n \text{ probability as } n\to\infty.$$

II 2. For some deterministic function $\sigma_t^2(\beta)$ such that $v_t(\beta) = \int_0^t \sigma^2(\beta)dF > 0$, $0<t\leqslant 1$, we have

$$\mathbb{V}_t^n(\beta) = \int_0^t |\mathbb{L}^n(\beta)|^2 dF^n \to v_t(\beta) \quad \text{in } P_\beta^n \text{ probability as } n\to\infty.$$

We shall show that the estimator $\hat{\beta}_n$, defined by (1.4) with $\Psi = \Psi^n$ and $\mathbb{N} = \mathbb{N}^n$ is consistent and asymptotically normal $N(0, 1/4v)$, $v = v_1(\beta)$. For this we need some additional conditions stipulated in the next two subsections.

5.2. Define $\frac{\partial}{\partial\beta}\Psi^n(\beta) = 2\{\mathrm{diag}\Psi^n(\beta)\}^{\frac{1}{2}}\frac{\partial}{\partial\beta}\sqrt{\Psi^n(\beta)}$ and

$$\frac{\partial}{\partial\beta}\ln\Psi^n(\beta) = 2\{\mathrm{diag}\Psi^n(\beta)\}^{-\frac{1}{2}}\frac{\partial}{\partial\beta}\sqrt{\Psi^n(\beta)}.$$

Obviously,

$$\int_0^t\{\frac{\partial}{\partial\beta}\ln\Psi^n(\beta)\}^T d\mathbb{A}^n = \int_0^t \mathbb{I}_n^T\frac{\partial}{\partial\beta}\Psi^n(\beta)d\bar{A}^n = 0. \tag{5.1}$$

Hence,

$$\int_0^t\{\frac{\partial}{\partial\beta}\ln\Psi^n(\beta)\}^T d\mathfrak{M}^n = \int_0^t\{\frac{\partial}{\partial\beta}\ln\Psi^n(\beta)\}^T d\mathbb{N}^n \tag{5.2}$$

and

$$\frac{1}{k_n}\int_0^t\{\frac{\partial}{\partial\beta}\ln\Psi^n(\beta)\}^T\mathrm{diag}\, d\mathbb{A}^n\{\frac{\partial}{\partial\beta}\ln\Psi^n(\beta)\} = 4\int_0^t|\mathbb{L}^n(\beta)|^2 dF^n\to 4v_t(\beta) \tag{5.3}$$

in P_β^n probability as $n\to\infty$ (see II.2).

Now we apply the Corollaries 4.1 and 4.2 to derive asymptotic normality of the integral in (5.2), taking account of (5.1) and (5.3), and then to establish the LAN property for the 'partial likelihood ratio' we define

$$Y_\beta^n(b) = \exp\int_0^1\{\ln\Psi^n(\beta_n)-\ln\Psi^n(\beta)\}^T d\mathbb{N}^n$$

with $\beta_n = \beta+b/2\sqrt{k_n v}$ where $b\in B_n = \mathfrak{B}-\beta/2\sqrt{k_n v}$ (see Remark 5.1 below). In addition we state the Lindeberg Condition III, under which the following Corollary 5.1 holds.

CONDITION III. If $(\bar{A}_t^n,\Psi_t^n)_\beta$ is the pair of the $(P_\beta^n,\mathfrak{F}_t^n)$-predictable characteristics, then for each $\epsilon>0$ Condition $\mathbb{H}'.1$ is satisfied with $\mathbb{H}^n = \frac{\partial}{\partial\beta}\ln\Psi^n$.

COROLLARY 5.1. *(i) If Conditions II and III are satisfied, then*

$$\mathfrak{L}(\frac{1}{\sqrt{k_n}}\int_0^1\{\frac{\partial}{\partial\beta}\ln\Psi^n(\beta)\}^T d\mathbb{N}^n|P_\beta^n)\Rightarrow N(0,4v)$$

(ii) Let the Conditions II and III be satisfied uniformly in $\beta\in K$, a compact subset of $\mathfrak{B}$. Then uniformly in $\beta\in K$ finite dimensional distributions of $Y_\beta^n(b)$ tend to finite dimensional distributions of

$$\exp\{b\zeta-\tfrac{1}{2}b^2\},\quad \mathfrak{L}(\zeta) = N(0,1).$$

REMARK 5.1. The term 'partial likelihood ratio' used above may be justified as follows (cf. Cox [5] or ANDERSEN AND GILL [1]):

Consider the situation described in the beginning of Subsection 4.2, and suppose that $\mathbb{N}^n$ possesses the pair $(\bar{A}_t^n,\Psi_t^n(\beta))$ of $(P^n,\mathfrak{F}_t^n)$-predictable characteristics and the pair $(\bar{A}_t^n,\Psi_t^n(\beta_n))$ of $(\underline{P}^n,\mathfrak{F}_t^n)$-predictable characteristics, such that the associated Hellinger process is bounded

$$\mathfrak{H}_t^n(P^n,\underline{P}^n)\equiv\mathfrak{H}_t^n = \int_0^t|\sqrt{\Psi^n(\beta_n)}-\sqrt{\Psi^n(\beta)}|^2 d\bar{A}^n<\infty\quad P^n\text{ a.s.}$$

Then by Theorem 3.1 and Remarks 3.3 and 3.4

$$\mathrm{E}(\mathrm{d}\underline{P}^n / \mathrm{d}P^n|\mathcal{F}_t) = \exp\int_0^t \{\ln\Psi^n(\beta_n) - \ln\Psi^n(\beta)\}^T \mathrm{d}\mathbb{N}^n$$

and

$$\mathrm{E}\{Y_\beta^n(b)|P^n\} = \mathrm{E}\{\exp\int_0^1 \{\ln\Psi^n(\beta_n) - \ln\Psi^n(\beta)\}^T \mathrm{d}\mathbb{N}^n | P^n\} = 1. \tag{5.4}$$

Here the likelihood ratio is called partial because the first characteristic is one and the same in the above considered pairs of $(P^n,\mathcal{F}_t^n)$- and $(P^n,\mathcal{F}_t^n)$- characteristics of $\mathbb{N}^n$ (According to this terminology, the estimator defined by (1.4) might be called the maximum partial likelihood estimator).

Note, in conclusion, that analogous considerations lead to

$$\mathrm{E}\{\sqrt{Y_\beta^n(b)}\exp(\tfrac{1}{2}\int_0^1 |\sqrt{\Psi^n(\beta_n)} - \sqrt{\Psi^n(\beta)}|^2 \mathrm{d}\bar{A}^{nc} +$$

$$+ \sum_{s\leqslant 1}(1-\Delta\bar{N}_s^n)\ln(1+\frac{1}{2}\frac{\Delta\bar{A}_s^n}{1-\Delta\bar{A}_s^n}|\sqrt{\Psi_s^n(\beta_n)} - \sqrt{\Psi_s^n(\beta)}|^2))|P^n\} = 1. \tag{5.5}$$

5.3. For establishing further properties of the partial likelihood ratio $Y_\beta^n(b)$ we require the following

CONDITION IV. If $(\bar{A}_t^n,\Psi_t^n)_\beta$ is the pair of the $(P_\beta^n,\mathcal{F}_t^n)$-predictable characteristics, with the parameter β belonging to a finite set $\mathcal{B}$, then the quantity

$$\delta_\beta^n(b_1,b_2) = \int_0^1 |\sqrt{\Psi^n(\beta_n^2)} - \sqrt{\Psi^n(\beta_n^1)}|^2 \overline{dA}^n, \tag{5.6}$$

where

$$\beta_n^i = \beta + b_i/2\sqrt{k_n v}, \quad b_i \in B_n = \mathcal{B} - \beta/2\sqrt{k_n v}, \; i=1,2,$$

obeys the following bounds: there are positive constants C_i (independent of n) such that for sufficiently large values of n and each $b_i \in B_n$, $i=1,2$, $b \in B_n$ and $\beta \in K$, a compact subset of $\mathcal{B}$,

IV. 1 $$P_\beta^n\{\delta_\beta^n(0,b) < C_0\frac{b^2}{1+b^2}\} \leqslant C_1 e^{-C_2 b^2};$$

cf. (5.6) with $b_1 = 0$, $b_2 = b$.

IV. 2 $$P_\beta^n \text{ a.s. } \delta_\beta^n(b_1,b_2) \leqslant C(b_2-b_1)^2$$

LEMMA 5.1. *Let Condition IV hold. Then there are positive constants c_i such that for each $b \in B_n$, $b_i \in B_n$, $i=1,2$ and $\beta \in K$, a compact subset of $\mathcal{B}$,*

(i) $E\{\sqrt{Y_\beta^n(b)}|P_\beta^n\} \leqslant c_1 e^{-c_2 b^2}$,

(ii) $E\{|\sqrt{Y_\beta^n(b_2)} - \sqrt{Y_\beta^n(b_1)}|^2 | P_\beta^n\} \leqslant c_2|b_2-b_1|^2$

PROOF. (i). By applying the inequality

$$E\{\sqrt{Y_\beta^n(b)}|P_\beta^n\} \leqslant E\{\sqrt{Y_\beta^n(b)}(\xi + I(\xi<1))|P_\beta^n\}$$

$$\leqslant E\{\sqrt{Y_\beta^n(b)}\xi|P_\beta^n\} + P_\beta^n(\xi<1)$$

(here (5.4) and the Schwarz inequality is used) to

$$\xi = \exp\tfrac{1}{2}\{\delta_\beta^n(0,b) - C_0\frac{b^2}{1+b^2}\}$$

and taking into account (5.5) and Condition IV we get

$$E\{\sqrt{Y^n_\beta(b)}|P^n_\beta\} \leqslant \exp\{-\frac{b^2}{2}(\frac{C_0}{1+b^2}-C)\}+C_1\exp(-C_2b^2)$$

This and the finiteness of $\mathfrak{B}$ imply (i).

(ii). By (5.4)

$$E\{|\sqrt{Y^n_\beta(b_2)}-\sqrt{Y^n_\beta(b_1)}|^2|P^n_\beta\} = 2(1-E\{\sqrt{Y^n_\beta(b_2)Y^n_\beta(b_1)}|P^n_\beta\})$$

and by (5.5) and Condition VI.2

$$1\leqslant E\{\sqrt{Y^n_\beta(b_2)Y^n_\beta(b_1)}\exp(\frac{1}{2}\sum_{s\leqslant 1}\frac{\Delta\bar{A}^n_s}{1-\Delta\bar{A}^n_s}|\sqrt{\Psi^n_s(\beta^2_n)}-\sqrt{\Psi^n_s(\beta^1_n)}|^2)|P^n_\beta\}\exp(\frac{1}{2}C(b_2-b_1)^2)$$

$$\leqslant\exp(\frac{1}{2}c(b_2-b_1)^2)E\{\sqrt{Y^n_\beta(b_2)Y^n_\beta(b_1)}|P^n_\beta\}$$

where the constant c arises by taking into consideration the arguments indicated in the footnote [1] on p. 42. Hence

$$E\{|\sqrt{Y^n_\beta(b_2)}-\sqrt{Y^n_\beta(b_1)}|^2|P^n_\beta\} \leqslant 2(1-e^{-\frac{1}{2}c(b_2-b_1)^2})\leqslant c(b_2-b_1)^2$$

□

Lemma 5.1 allows us to apply the result of SIEDERS AND DZHAPARIDZE [24] mentioned in subsection 2.1, a version of the 'large deviations' Theorem 5.1 of IBRAGIMOV AND HAS'MINSKII [11], in order to draw inference about the estimator $\hat{\beta}_n$ for the parameter β in the spirit of the above mentioned book (see the next subsection). For this end we have resticted our attention to the scalar parameter case in which the property of the partial likelihood $Y^n_\beta(b)$ stated in assertion (ii) of Lemma 5.1 turns out to be sufficient (in fact, if β were a p-vector valued parameter then one should guarantee the property (5.8) below with $2m > p$).

REMARK 5.2. Condition IV.1 is a consequence of the identifiability condition: for each $\delta > 0$

$$\inf_{|h|\geqslant\delta}\int_0^1|\sqrt{\Psi^n(\beta+h)}-\sqrt{\Psi^n(\beta)}|^2\mathrm{d}\bar{A}^n > 0,\quad P^n_\beta\text{a.s.} \tag{5.7}$$

provided that

$$\delta^n_\beta(0,b)\to\frac{b^2}{4}\quad\text{in } P^n_\beta \text{ probability as } n\to\infty \tag{5.8}$$

Indeed (5.7) and (5.8) allow us to choose a constant C_0 such that for sufficiently large values of n and each $\epsilon > 0$

$$P^n_\beta\{\delta^n_\beta(0,b) < C_0\frac{b^2}{1+b^2}\} \leqslant \epsilon$$

(cf. IBRAGIMOV AND HAS'MINSKII [11], p. 82), and this implies Condition IV.1.

Observe that in the case that the Newton-Leibniz formula holds

$$\sqrt{\Psi^n(\beta+h)}-\sqrt{\Psi^n(\beta)} = h\int_0^1\mathbb{L}^n(\beta+sh)\mathrm{d}s \tag{5.9}$$

(cf. IBRAGIMOV AND HAS'MINSKII [11], p. 66), the condition (5.8) can be viewed as a strengthening of Condition II.2, because (5.8) can be rewritten as

$$\int_0^1 \Big| \int_0^1 \mathbb{L}^n(\beta + sh / \sqrt{k_n}) \mathrm{d}s \Big|^2 \mathrm{d}F^n \to v \tag{5.10}$$

in P_β^n probability as $n \to \infty$, for each h and not only for $h = 0$ as in Condition II.2.

By the same arguments, if

$$\sup_b \int_0^1 |\mathbb{L}^n(\beta_n)|^2 \mathrm{d}F^n < \infty \quad P_\beta^n \text{ a.s.}, \tag{5.11}$$

then Condition IV.2 is also satisfied.

5.4. As a consequence of Lemma 5.1 and Theorem 2.1 in SIEDERS AND DZHAPARIDZE [24] we have

PROPOSITION 5.1. *Under Condition IV there are certain positive constants C_0 and c_0 such that the estimator $\hat{\beta}_n$ is consistent: $\hat{\beta}_n \to \beta$ in P_β^n probability, and for sufficiently large values of n and H*

$$\sup_{\beta \in K} P_\beta^n \{ |2\sqrt{k_n v}(\hat{\beta}_n - \beta)| > H \} \leqslant C_0 e^{-c_0 H^2}$$

In view of the assertion (ii) of Corollary 5.1 and Lemma 4.1, we can make use of the Theorems I.10.1 and III.1.2 in IBRAGIMOV AND HAS'MINSKII [11]. The result can be formulated as

PROPOSITION 5.2. *If the Conditions II-IV hold uniformly in $\beta \in K$, then we have asymptotically as $n \to \infty$*

$$P_\beta^n \{ |2\sqrt{k_n v}(\hat{\beta}_n - \beta) - \frac{1}{2\sqrt{k_n v}} \int_0^1 \{ \frac{\partial}{\partial \beta} \ln \Psi^n(\beta) \}^T d\mathbb{N}^n | > \delta \} \to 0, \text{ for each } \delta > 0$$

$$\mathcal{L}\{ 2\sqrt{k_n v}(\hat{\beta}_n - \beta) | P_\beta^n \} \Rightarrow N(0,1)$$

and

$$E\{ |2\sqrt{k_n v}(\hat{\beta}_n - \beta)|^J | P_\beta^n \} \to \frac{2^{\frac{J}{2}}}{\sqrt{\pi}} \Gamma(\frac{J+1}{2})$$

uniformly in $\beta \in K$.

5.5. In the remainder of this paper we restrict our attention to the special case desrcibed in Example 1.1 in which the process $\mathbb{N}^n$ is of the Poisson type with the compensator (1.1) where $\Phi = \Phi^n$ satisfies the conditions stipulated below.

CONDITION ΦI. There is a sequence of r_n-vector valued $\mathcal{F}_t^n$-predictable processes, continuously dependent on β, say $(\partial/\partial\beta)\sqrt{\Phi^n(\beta)}$, $n=1,2,\ldots$, such that for each $\alpha\in\mathcal{A}$ and $\beta\in\mathcal{B}$ the following holds:

ΦI.1. For each b such that $\underline{\beta}^n=\beta+b/\sqrt{k_n}\in\mathcal{B}$ eventually,

$$\int_0^1 |\sqrt{\Phi^n(\underline{\beta}^n)}-\sqrt{\Phi^n(\beta)}-\frac{b}{\sqrt{k_n}}\frac{\partial}{\partial\beta}\sqrt{\Phi^n(\beta)}|^2 d\alpha\to 0 \text{ in } P^n_{\alpha,\beta} \text{ probability as } n\to\infty;$$

ΦI.2. For some deterministic function $\rho_t^2(\beta)$ such that $w_t(\alpha,\beta)=\int_0^t \rho_s^2(\beta)d\alpha_s>0, 0<t\leqslant 1$, and for each t, $0\leqslant t\leqslant 1$, $\alpha\in\mathcal{A}$, $\beta\in\mathcal{B}$

$$\mathcal{W}_t^n(\alpha,\beta)=\frac{1}{k_n}\int_0^t |\frac{\partial}{\partial\beta}\sqrt{\Phi_s^n(\beta)}|^2 d\alpha_s\to w_t(\alpha,\beta) \text{ in } P^n_{\alpha,\beta} \text{ probability as } n\to\infty.$$

CONDITION ΦII. There is a positive bounded deterministic function $\phi_t(\beta)$ (uniformly in t and β $m<\phi_t(\beta)<M$ where $0<m<M<\infty$) having continuous bounded derivative in β, such that for each $\alpha\in\mathcal{A}$, $\beta\in\mathcal{B}$ and $0\leqslant t\leqslant 1$

ΦII.1. $F_t^n(\alpha,\beta)=\int_0^t \frac{1}{k_n}\bar{\Phi}_s^n(\beta)d\alpha_s\to F_t(\alpha,\beta)$ in $P^n_{\alpha,\beta}$ probability as $n\to\infty$ where $F_t=F_t(\alpha,\beta)=\int_0^t \phi_s(\beta)d\alpha_s>0$ (cf. Condition I).

ΦII.2. For $(\partial/\partial\beta)\bar{\Phi}^n = \mathbb{I}_{r_n}^T(\partial/\partial\beta)\Phi^n = 2\mathbb{I}_{r_n}^T\{\mathrm{diag}\Phi^n\}^{1/2}(\partial/\partial\beta)\sqrt{\Phi^n}$

$$\frac{1}{k_n}\int_0^t \frac{\partial}{\partial\beta}\bar{\Phi}_s^n(\beta)d\alpha_s\to\int_0^t \frac{\partial}{\partial\beta}\phi_s(\beta)d\alpha_s<\infty \text{ in } P^n_{\alpha,\beta} \text{ probability as } n\to\infty.$$

REMARK 5.4. It is easily seen that these conditions imply Condition II for entries as in (1.2). Besides, by the Conditions II.2, ΦI.2, ΦII.1, ΦII.2 and (5.3)

$$\mathcal{V}_t(\alpha,\beta)=\mathcal{W}_t(\alpha,\beta)-\frac{1}{k_n}\int_0^t |\frac{\partial}{\partial\beta}\sqrt{\bar{\Phi}_s^n}|^2 d\alpha_s\to w_t(\alpha,\beta)-\int_0^t |\frac{\partial}{\partial\beta}\sqrt{\phi_s}|^2 d\alpha_s = v_t(\alpha,\beta)$$

in $P^n_{\alpha,\beta}$ probability as $\to\infty$.

It is easily seen that Condition III follows from ΦII and the following Lindeberg condition:

CONDITION ΦIII. For each $\epsilon>0$ Condition $\mathbb{H}'.1$ of Section 4 is satisfied with $\mathbb{H}^n=\frac{\partial}{\partial\beta}\ln\Phi^n\equiv 2\{\mathrm{diag}\Phi^n\}^{-1/2}\frac{\partial}{\partial\beta}\sqrt{\Phi^n}$.

5.6. Here we apply the results of Propositions 5.1 and 5.2 (as well as that of Corollary 6.1 below) to the special case of the Cox regression model: see (1.3) with the covariate processes $Z_t^{in}, i=1,...,r_n$ and the censoring factors $Y_t^{in}, i=1,...,r_n$ taking values 0 or 1 (so N_t^{in} only jumps when $Y_t^{in}=1$). To arrive at these statements concerning the properties of the estimator β_n we adopt the assumptions of ANDERSEN AND GILL [1], strengthening for simplicity Condition C on p. 1105 (in fact its scalar version given on p. 1110) to the following condition:

$$\sup_{i,t}|Z_t^{in}|Y_t^{in}<\infty \quad P_{\alpha,\beta}^n \text{ a.s.} \tag{5.14}$$

For this end we shall show that all the conditions stipulated above are satisfied. By (5.14) the Lindeberg condition ΦIII is completely empty, and so is the condition (5.11).

Since in this particular case

$$\left|\frac{\partial}{\partial\beta}\sqrt{\Phi^n}\right|^2 = \frac{1}{4}\frac{\partial^2}{\partial\beta^2}\bar{\Phi}^n, \tag{5.15}$$

Condition ΦI.2 follows from Condition B in ANDERSEN AND GILL [1], p. 1105. The latter condition, along with Condition D in [1], ensures also ΦII, with

$$v(\alpha,\beta) = \frac{1}{4}\int_0^1\{(\partial^2/\partial\beta^2)\phi(\beta) - \frac{[(\partial/\partial\beta)\phi(\beta)]^2}{\phi(\beta)}\}d\alpha > 0.$$

Further, applying (5.14) and (5.15), the Newton-Leibniz formula (cf. (5.9)), and the above mentioned Conditions B and D, one easily verifies ΦI.1. Finally, by the same arguments one gets (5.8) which gives Condition IV.1 because here the identifiability (5.7) takes place.

6. ASYMPTOTIC OPTIMALITY.

6.1. Suppose that the process $\mathbb{N}^n$ is of the Poisson type and retain the Conditions ΦI-ΦIII stipulated in Subsection 5.5. Let the probability measures P^n and $\underline{P}^n$ be defined on $(\Omega^n,\mathcal{F}^n)$ as in Subsection 2.2. Then (see Remark 3.4)

$$\frac{d\underline{P}^n}{dP^n} = \exp\{\int_0^1(\ln\underline{\Phi}_s^n(\underline{\beta}^n)-\ln\Phi_s^n(\beta))^T d\mathbb{N}_s^n + 2\int_0^1\ln\frac{d\underline{\alpha}_s^n}{d\alpha_s^n}d\bar{N}_s^n \tag{6.1}$$

$$-\int_0^1\bar{\underline{\Phi}}_s^n(\underline{\beta}^n)d\underline{\alpha}_s^n + \int_0^1\bar{\Phi}_s^n(\beta)d\alpha_s\}$$

We shall now apply Corollary 4.2 to show the LAN of the $\{P_{\alpha,\beta}^n, \alpha\in\mathcal{A}, \beta\in\mathcal{B}\}$ in the sense of

DEFINITION 6.1. The family $\{P_{\alpha,\beta}^n, \alpha\in\mathcal{A}, \beta\in\mathcal{B}\}$ is called Locally Asymptotically Normal (LAN) at the 'point' $\alpha\in\mathcal{A}$, $\beta\in\mathcal{B}$ if for each $b\in R^1$ and each bounded $a\in L^2(dF)$ such that $\underline{\alpha}^n\in\mathcal{A}, \underline{\beta}^n\in\mathcal{B}$ eventually, there is a sequence of asymptotically normal variables $\delta_{\alpha,\beta}^n(a,b)$, $n=1,2,\ldots$:

$$\mathcal{L}\{\delta_{\alpha,\beta}^n(a,b)|P_{\alpha,\beta}^n\}\Rightarrow N(0,g_{\alpha,\beta}(a,b)) \text{ as } n\to\infty$$

with $g_{\alpha,\beta}(a,b)>0$ for which $d\underline{P}^n/dP^n = \exp\{\delta_{\alpha,\beta}^n(a,b)-\frac{1}{2}g_{\alpha,\beta}(a,b) + \eta_{\alpha,\beta}^n(a,b)\}$ where $\eta_{\alpha,\beta}^n(a,b)\to 0$ in $P_{\alpha,\beta}^n$ probability as $n\to\infty$.

Note first that if

$$\mathbb{U}_t^n = \text{col}\{\sqrt{\Phi_t^{in}(\underline{\beta}^n)/\Phi_t^{in}(\beta)}\sqrt{d\underline{\alpha}_t^n/d\alpha_t}-1,\ i=1,...,r_n\},$$

then (4.29) is satisfied by $\mathbb{S}_t^n = \mathbb{S}_{t,\beta}^n(a,b) = \frac{1}{2}b(\partial/\partial\beta)\ln\Phi_t^n(\beta) + a_t\mathbb{1}_{r_n}$, for which

$$\frac{1}{k_n}\int_0^t \mathbb{S}^{nT}\mathrm{diag}\,d\mathbb{A}^n\mathbb{S}^n \to b^2 w_t(\alpha,\beta) + 2b\int_0^t \frac{(\partial/\partial\beta)\sqrt{\phi_s(\beta)}}{\sqrt{\phi_s(\beta)}} a_s dF_s(\alpha,\beta) + \int_0^t a_s^2 dF_s(\alpha,\beta) \qquad (6.2)$$

in P^n probability as $n\to\infty$, by the Conditions ΦI and ΦII.

PROPOSITION 6.1. *Under the Conditions* ΦI-ΦIII *the family* $\{P^n_{\alpha,\beta}, \alpha\in\mathcal{A}, \beta\in\mathcal{B}\}$ *is LAN at the 'point'* $\alpha\in\mathcal{A}$, $\beta\in\mathcal{B}$ *for* $\delta^n_{\alpha,\beta}(a,b) = \frac{1}{\sqrt{k_n}}\int_0^1 \mathbb{S}^n_\beta(a,b)^T d(\mathbb{N}^n - \mathbb{A}^n(\alpha,\beta))$ *and* $g_{\alpha,\beta}(a,b)$ *then equals 4 times the right-hand side of (6.2) evaluated at* $t=1$.

6.2. Suppose that the underling model confines 'the directions' a to a linear subspace $\mathbf{A}\in L^2(dF)$, and let $\{\beta^n_{RA}\}$ be a class of $R^1\times\mathbf{A}$-regular estimators for β, which includes a subclass of regular estimators $\{\beta^n_R\}\subset\{\beta^n_{RA}\}$ (see Subsection 2.3). Then, by Hájek's convolution theorem (BEGUN et al. [2], Theorem 3.1), we have

PROPOSITION 6.2. *Let the Condition* ΦI-ΦIII *hold. Then*

(i) $$\mathcal{L}\{\sqrt{k_n}(\beta^n_{RA}-\beta^n)|P^n\}\Rightarrow G^0_{RA} = N(0,\mathcal{I}^{-1}_{\alpha,\beta})*G^1_{RA} \qquad (6.3)$$

as $n\to\infty$ *with some distribution law* G^1_{RA}, *where* $\mathcal{I}_{\alpha,\beta} = 4\{w_1(\alpha,\beta) - \int_0^1 \pi_s^2(\alpha,\beta)dF_s(\alpha,\beta)\}$, $\pi_s(\alpha,\beta)$ *being the projection of* $(\partial/\partial\beta)\sqrt{\phi_t(\beta)}/\sqrt{\phi_t(\beta)}$ *into* $\mathbf{A}$, *that is, it satisfies the equation*

$$\int_0^1 \left\{\frac{(\partial/\partial\beta)\sqrt{\phi_s(\beta)}}{\sqrt{\phi_s(\beta)}} - \pi_s(\alpha,\beta)\right\} a_s dF_s(\alpha,\beta) = 0$$

for each $a_s\in\mathbf{A}$.

(ii) $$\mathcal{L}\{\sqrt{k_n}(\beta^n_R-\beta^n)|P^n\}\Rightarrow G^0_R = N(0,1/4v_1(\alpha,\beta))*G^1_R \qquad (6.4)$$

as $n\to\infty$ *with some distribution law* G^1_R *(see Remark 5.4), uniformly for each* $|b|<c$ *whatever* $c>0$, *and each bounded* $a\in L^2(dF)$

REMARK 6.1. For the 'least favorable' direction $a_t = -b\pi_t(\beta)$ the quantity $g_{\alpha,\beta}(a,b)$ coincides with $b^2\mathcal{I}_{\alpha,\beta}$, $\mathcal{I}_{\alpha,\beta}$ being Fisher's information for β (BEGUN et al. [2], Section 3).

REMARK 6.2. Evidently, $\mathcal{I}_{\alpha,\beta}\geqslant 4v_1(\alpha,\beta)$ with equality iff $(\partial/\partial\beta)\sqrt{\phi_t(\beta)}/\sqrt{\phi_t(\beta)}\in\mathbf{A}$ (see Remark 5.4). Having the limiting distribution of $\sqrt{k_n}(\hat\beta_n-\beta)$ under P^n (Proposition 5.2), one can apply the usual contiguity arguments (allowed by Proposition (6.1)) to arrive at the formula (6.4) for $\beta^n_R = \hat\beta_n$ with G^1_R that degenerates at 0. These considerations can be summarized as the following statement on the optimality properties of the estimator $\hat\beta_n$.

THEOREM 6.1. *Under the conditions stipulated above, the estimator* $\hat\beta_n$ *is regular and* $\mathcal{L}\{\sqrt{k_n}(\hat\beta_n-\beta^n)|P^n\}\Rightarrow N(0,1/4v_1(\alpha,\beta))$ *for each* $b\in R^1$ *and each bounded* $a\in L^2(dF)$ *determining* α^n,β^n *and* P^n *as in Subsection 2.2.*

The estimator $\hat\beta_n$ is the best among $\{\beta^n_R\}$ in the sense that no regular estimator can have its limiting distribution less spread then $\hat\beta_n$. Besides, iff $(\partial/\partial\beta)\sqrt{\phi_t(\beta)}/\sqrt{\phi_t(\beta)}\in\mathbf{A}$, then it is the best among $\{\beta^n_{RA}\}\supset\{\beta^n_R\}$ in the same sense. Finally, observe that Proposition 6.2 and Theorem 11.2 in IBRAGIMOV AND HAS'MINSKII [11], Chapter II, allow us to obtain the lower bounds for the risk of $R^1\times\mathbf{A}$-regular and regular estimators and, consequently, to give yet another characterization of the optimality of $\hat\beta_n$. Namely, the following corollary holds:

COROLLARY 6.1 *Let the conditions stipulated above be satisfied. Let* $w(x)\geqslant 0$ $x\in R^1$ *be a continuous even loss function. Then for fixed* $\alpha\in\mathcal{A},\beta\in\mathcal{B}$

$$\liminf_{n\to\infty} E\{w(\zeta^n) \mid P^n_{\alpha,\beta}\}\geqslant \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} w(x)e^{-\frac{1}{2}x^2}\,dx$$

where $\zeta^n=2\sqrt{k_n v}(\beta^n_R-\beta)$. *The same inequality holds also for* $\zeta^n=\sqrt{k_n\vartheta}(\beta^n_{RA}-\beta)$. *In particular*

$$\liminf_{n\to\infty} k_n var\{\beta^n_R \mid P^n_{\alpha,\beta}\}\geqslant(4\pi)^{-1}\geqslant\vartheta^{-1},$$

and

$$\liminf_{n\to\infty} k_n var\{\beta^n_{RA} \mid P^n_{\alpha,\beta}\}\geqslant\vartheta^{-1}$$

If, in addition, w allows a polynominal majorant, then by the last assertion of Proposition 4.2

$$\lim_{n\to\infty} E\{w(2\sqrt{k_n v}(\hat{\beta}_n-\beta)) \mid P^n_{\alpha,\beta}\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} w(x)e^{-\frac{1}{2}x^2}\,dx,$$

hence $\hat{\beta}_n$ *attains the lower bound for the risks of regular estimators. Besides, iff* $(\partial/\partial\beta)\sqrt{\phi_t(\beta)}\,/\,\sqrt{\phi_t(\beta)}\in\mathbf{A}$, *then* $\vartheta=4v$ *and* $\hat{\beta}_n$ *attains the lower bound also for the risks of* $R^1\times\mathbf{A}$*-regular estimators.*

REFERENCES.

1. P. K. ANDERSEN AND R. D. GILL (1982), Cox's regression model for counting processes: a large sample study, *Ann. Statist. 10,* 1100-1120.
2. J.M. BEGUN, W.J. HALL, W.M. HUANG AND J.A. WELLNER (1983), Information and asymptotic efficiency in parametric and non-parametric models, *Ann. Statist. 11,* 432-452.
3. Ø. BORGAN (1984), Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scand. J. Statist. 11,* 1-16.
4. P. BRÉMAUD (1981), *Point Processes and Queues. Martingale Dynamics.* Springer-Verlag. New York.
5. D. R. COX (1975), Partial likelihood. *Biometrika 62,* 269-276.
6. K. O. DZHAPARIDZE (1986), *On LAN for Counting Processes,* Report MS-R8606, Centrum voor Wiskunde en Informatica, Amsterdam.
7. B. EFRON (1977), The efficiency of Cox's likelihood function for censored data *J. Amer. Statist. Assoc. 72,* 557-565.
8. R. D. GILL AND S. JOHANSEN (1986), *Product integrals and counting processes,* Report MS-R86xx, Centrum voor Wiskunde en Informatica, Amsterdam.
9. P. E. GREENWOOD AND A. N. SHIRYAYEV (1985), Continuity and the Statistical Invariance Principle. *Stochastic Monographs. Vol. 1,* Gordon and Breach, New York.
10. N. L. HJORT (1984), Bayes estimators and asymptotic efficiency in parametric counting process models. Research report., Norw. Comp. Centre, Oslo (submitted to *Scand. J. Statist).*
11. I. A. IBRAGIMOV AND A. N. HAS'MINSKII (1981), *Statistical Estimation. Asymptotic Theory.* Springer-Verlag. New York.
12. B. D. S. JARUPSKIN (1983), *Maximum likelihood and related estimation methods in point processes and point process systems.* Ph. D., Dept. Statist., University of California, Berkeley.
13. YU. M. KABANOV, R. S. LIPTSER AND A. N. SHIRYAYEV (1975), Criteria of absolute continuity of measures corresponding to multivariate point processes. Proc. Third Japan-USSR Sympos. Probability Theory (Tashkent, 1975), *Lecture Notes in Math., vol. 550,* Springer-Verlag, New York.
14. YU. M. KABANOV, R. S. LIPTSER AND A. N. SHIRYAYEV (1979), Absolute continuity and singularity of locally absolute continuous probability distributions. I *Math. USSR Sbornik. 35,* 631-679.
15. YU. M. KABANOV, R. S. LIPTSER AND A. N. SHIRYAYEV (1980), Absolute continuity and

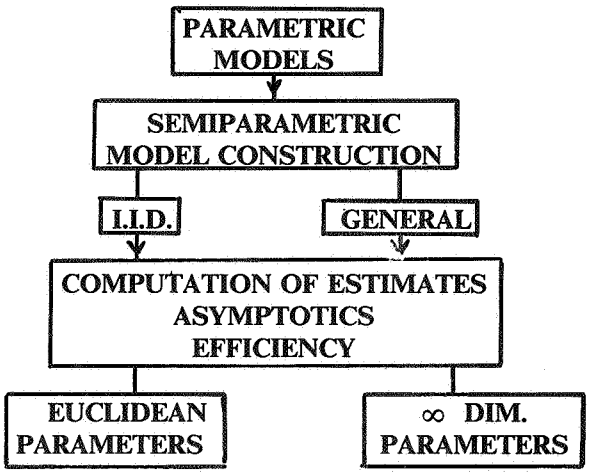
singularity of locally absolute continuous probability distributions. II *Math. USSR Sbornik. 36,* 31-58.

16. YU. M. KABANOV, R. S. LIPTSER AND A.N. SHIRYAYEV (1980), Some limit theorems for simple point processes (a martingale approach). *Stochastics. 3,* 203-216.
17. L. LE CAM (1969), *Theorie Asymptotique de la Decision Statistique,* Les Presses de l'Universite de Montreal. Montreal.
18. R. S. LIPTSER AND A. N. SHIRYAYEV (1978), *Statistics of Random Processes II. Applications.* Springer-Verlag. New York.
19. R. S. LIPTSER AND A. N. SHIRYAYEV (1980), A functional central limit theorem for semimartingales. *Theory Probab. Appl. 25,* 667-688.
20. D. L. MCLEISH (1978), An extended martingale principle. *Ann. Probab. B6,* 144-150.
21. P. W. MILLAR (1983), The minimax principle in asymptotic statistical theory, *Proc. Ecole d'Ete St. Flour, Lecture Notes in Math. 976,* 75-265, Springer-Verlag, Berlin.
22. R. L. PRENTICE AND S. G. SELF (1983), Asymptotic distribution theory for Cox-type regression models with general relative risk form. *Ann. Statist. 11,* 804-813.
23. A. N. SHIRYAYEV (1981) Martingales: Recent Developments, Results and Applications. *Int. Statist. Review. 49,* 199-233.
24. A. SIEDERS AND K. O. DZHAPARIDZE (1986), A large deviation result for parameter estimators and its applications to nonlinear regression analysis, (submitted to *Annals of Statistics).*

Discussion of papers on Semiparametric Models

P.J. Bickel
Department of Mathematical Statistics,
University of California at Berkeley,
Berkeley Ca. 94720,
U.S.A.

The three interesting papers we have just heard illustrate the broad scope of this topic. This very breadth makes discussion difficult and I apologize to the authors and audience in advance for sins of omission. Let me begin with a schematic overview of what I see as the main areas of activity in estimation in semiparametric models and position today's papers in this schema.



Wellner gives us an overview of all aspects of this diagram pertaining to estimation of Euclidean parameters in the i.i.d. case. *Dzhaparidze*, generalizing Andersen and Gill's [1] work, establishes, in a counting process framework, the asymptotic optimality properties of partial likelihood estimates for the parameters of what might be called the general proportional hazards model. *Clayton and Cuzick* (CC), to whose paper I'll devote most of my discussion, introduce the semiparametric Pareto model in the i.i.d. case (with censoring). This model, a generalization of the Cox proportional hazards model, is obtained here by extending the parametric Pareto model even as Cox extended the classical exponential failure time model. The mechanism, in the uncensored case, can be described in Wellner's terminology, which I will also use in the sequel, as follows.

Given a parametric model $\{P_\theta : \theta \in \Theta\}$ for $X=(Z,T)$, generate a semi-parametric model $\{P_{(\theta, G)} : \theta \in \Theta, G^{-1} = h \in \mathcal{H}\}$ in which $X=(Z, h(T))$ where $(Z, T) \simeq P_\theta$ and $\mathcal{H}$ is the large

group of monotone increasing transformations. Then $(Z_1, \ldots, Z_n, R_1, \ldots, R_n)$, where $R_1, \ldots, R_n$ are the ranks of $h(T_1), \ldots, h(T_n)$, is a maximal invariant under $\mathcal{H}$. It is reasonable to expect that efficient estimates of θ (if θ is identifiable) can be based on the marginal likelihood of $(Z_1, \ldots, Z_n, R_1, \ldots, R_n)$.

CC suggest, on heuristic grounds, an algorithm which should converge to the maximum partial likelihood estimate of θ and then suggest an estimate for h also on likelihood grounds. Although they deal extensively with computation of the estimates and estimates of their variances, consistency, the validity of the normal approximation, consistency of the estimates of variance, and efficiency are not dealt with, at least in this paper. I'd like to sketch a programme for establishing these properties for these estimates or at least one step approximations to them. Let me begin by reviewing essentially what calculations are needed to establish the asymptotic behaviour of estimates based on maximizing 'pieces' of the likelihood which do not depend on the infinite dimensional part G of the parameter (θ, G).

Suppose we are given $X_1, \ldots, X_n$, i.i.d. with distribution $P_{(\theta,G)}$ and corresponding density $p(\cdot, \theta, G)$. The gradient of the partial or marginal likelihood, $w_n(\theta, X_1, \ldots, X_n)$, and a corresponding partial or marginal likelihood satisfies

$$E_{(\theta,G)} w_n(\theta) = 0 \quad \forall\, G.$$

The maximum 'likelihood' estimator $\hat{\theta}$ solves the equation $w_n(\hat{\theta}) = 0$. Now $\hat{\theta}$ is asymptotically normal if:

$$\hat{\theta} = \theta - \frac{w_n(\theta)}{w_n'(\theta)} + o_P(n^{-1/2}), \tag{i}$$

$$n^{1/2} w_n(\theta) \to^{\mathcal{L}} \mathcal{N}(0, \sigma^2(\theta, G)), \tag{ii}$$

$$w_n' \to^{P} c(\theta, G) \neq 0. \tag{iii}$$

In the i.i.d. case we usually have

$$w_n(\theta) = -n^{-1} \sum_{i=1}^{n} \gamma(X_i, \theta, G) + o_P(n^{-1/2}) \tag{iv}$$

where

$$\int \gamma^2(x, \theta, G) dP < \infty, \quad \int \gamma(x, \theta, G) dP = 0$$

and

$$w_n'(\theta) = \int \gamma_\theta(x, \theta, G) dP + o_P(1).$$

Then, under uniformity conditions,

$$\hat{\theta} = \theta + n^{-1} \sum_{i=1}^{n} \tilde{\gamma}(X_i, \theta, G) + o_P(n^{-1/2})$$

where

$$\tilde{\gamma} = \gamma / E(\gamma \ell_\theta(X_1))$$

and

$$\ell_\theta = \frac{\partial}{\partial \theta} \log p(\cdot, \theta, G).$$

Here are some properties of γ:

$$\gamma \perp \dot{\mathcal{P}}_G = \{\frac{\partial}{\partial \eta} \log p(X_1, \theta, G_\eta)|_{\eta=0}, G_0 = G\}, \tag{a}$$

$$\hat{\theta} \text{ is efficient} \Leftrightarrow \gamma = \ell_\theta - \Pi(\ell_\theta \mid \dot{\mathcal{P}}_G) \Leftrightarrow \gamma - \ell_\theta \in \dot{\mathcal{P}}_G. \tag{b}$$

Note that property (i) is easier to establish for a 1-step approximation to a root of $w_n(\cdot)$ starting from a $\sqrt{n}$-consistent estimate $\tilde{\theta}$ of θ. We return to the Cox and Pareto models in the context of special transformation models:

$$(Z, T) \simeq (Z, h(T')) \simeq P_{(\theta, G)},\ G = h^{-1};\ (Z, T') \simeq P_\theta.$$

Hence if P_θ has density $f_\theta(z, t')$ the two models are specified by:

$$f_\theta(z, t') = \nu(z, \theta)e^{-t'\nu(z, \theta)}q_0(z) \tag{Cox}$$

where q_0 is the marginal density of Z and conventionally $\nu(z, \theta) = e^{\theta^T z}$;

$$f_\theta(z, t') = \nu(1+\nu\gamma t')^{-(1+\frac{1}{\gamma})}q_0(z). \tag{Pareto}$$

For simplicity suppose θ is real (i.e. one-dimensional). Then the marginal likelihood argument leads us to take

$$w_n(\theta, \underline{r}, \underline{z}) = n^{-1}\frac{\partial}{\partial\theta}\log P_\theta[\underline{R} = \underline{r} \mid \underline{Z} = \underline{z}]$$

where $\underline{R} = (R_1, \ldots, R_n)$ is the vector of ranks. Then

$$w_n(\theta) = n^{-1}\sum_{i=1}^{n} E_\theta\{\ell_\theta(Z_i, T_i) \mid \underline{Z}, \underline{R}\}$$

where $\ell_\theta = (\partial / \partial\theta)\log f_\theta$. In the Cox model, $\ell_\theta = z(1-\nu t)$,

$$w_n(\theta) = n^{-1}\sum_{i=1}^{n} Z_{D_i}\left[1 - \nu_i\sum_{j=1}^{i}\frac{1}{\sum_{l=j}^{n}\nu_l}\right],$$

where $T_{(i)} = T_{D_i}$ and $\nu_l = \nu(z_{D_l}, \theta)$. We specialize to the two sample problem, $P[Z = 1] = \pi = 1 - P[Z = 0]$. Let F_n be the empirical d.f. of the T_i's, $M = \Sigma_{i=1}^n Z_i$, $\hat{\pi} = M / n$, $F_{0n} = (n-M)^{-1}\Sigma_{i=1}^n(1-Z_i)I(T_i \leqslant \cdot)$, $F_{1n} = M^{-1}\Sigma_{i=1}^n Z_i I(T_i \leqslant \cdot)$,

$$A(s) = \int_0^s \left[\int_\nu^1 [e^\theta\hat{\pi}\mathrm{d}F_{1n}F_n^{-1}(t) + (1-\hat{\pi})\mathrm{d}F_{0n}F_n^{-1}(t)]\right]^{-1} \mathrm{d}\nu.$$

Then

$$w_n(\theta) = \hat{\pi}\int_0^\infty (1 - e^\theta A(F_n(x)))\mathrm{d}F_{1n}(x).$$

Standard von Mises calculations (see TSIATIS [5] for rigorous treatment) give γ and $\ell_\theta - \gamma \in \dot{\mathcal{P}}_G$.

In the Pareto two sample model, the CC heuristics lead to the following $\tilde{w}_n$ as an approximation to w_n. Let

$$v(t) = 1 + e^\theta\frac{f'}{f}(e^\theta t)t,\ f(t) = (1+\gamma t)^{-(1+\frac{1}{\gamma})},$$

order $v(T_1), \ldots, v(T_n)$ as $V_{(1)} < \ldots < V_{(n)}$, and let $F_j = \mathcal{L}(v(T_1) \mid Z_1 = j)$, $F_j' = f_j$. Also let $\delta_j = I(Z_{\tilde{D}_j} = 1)$, $V_{(j)} = v(T_{\tilde{D}_j})$, and $b_j(x) = -\frac{f_j'}{f_j}(x)$ (here, θ is hidden). Then

$$w_n(\theta) \approx \tilde{w}_n(\theta) = \hat{\pi}\int_0^\infty (1 - e^\theta B(F_n(x)))\mathrm{d}F_{1n}(x),$$

where

$$B(s) = d + \int_0^s \left[\int_v^1 \left\{ \hat{\pi} b_1(B(t)) \mathrm{d}F_{1n} F_n^{-1}(t) + (1-\hat{\pi}) b_0(B(t)) \mathrm{d}F_{0n} F_n^{-1}(t) \right\} + c \right]^{-1} \mathrm{d}v.$$

This is the same formula as for the Cox model but B is explicit in that case since $c = d = 0$, $b_0 \equiv 1$, $b_1 \equiv e^{\theta}$.

The formal von Mises calculation here is messier but should also lead to $\ell_\theta - \gamma \in \dot{\mathcal{P}}_G$. However, there are major additional technical difficulties involving the establishment of the existence of B (which as CC point out has to be defined more carefully) and checking (iv), (iii) and then (i) but at the moment these difficulties do not seem insurmountable even if we add censoring. Note that if we can establish (iv) then by contiguity calculations the efficiency of the estimate based on $\tilde{w}_n$ will imply, modulo a uniformity argument, that $[-\tilde{w}'_n(\theta)]^{-1}$ is a consistent estimate of the asymptotic variance of $\sqrt{n}\hat{\theta}$. One may ask why one doesn't rush to apply the methods mentioned by Wellner. The difficulty is that, at least so far, we do not know how to calculate the projection on $\dot{\mathcal{P}}_G$ explicitly.

I conclude with some brief comments on the Wellner and Dzhaparidze papers. On the whole I found Wellner's classification of models very reasonable. However, I believe the distinction between semiparametric and nonparametric models is more of description than of size. All models (including parametric ones!) can be described nonparametrically as a set of probability distributions. Semiparametric models are ones which are 'naturally (smoothly)' describable by means of a mixed Euclidean and infinite dimensional parametrization even as parametric models are 'smoothly' describable by means of a Euclidean parametrization.

Although estimation methods 1 and 2 have proved successful in a variety of contexts they are limited by the need for ad hoc construction of preliminary estimates and calculation of projections. Nonparametric maximum likelihood (NPML), Hellinger and Bayes methods (with and without restricting to a sieve) do not have these limitations. On the other hand NPMLE's can be nonexistent as in regression models or consistently inconsistent as in the IFRA-distributions model of BOYLES et al. [3] or in the earlier example referred to there of starshaped distributions. Yet NPML works in the Cox (BAILEY [2] or JOHANSEN [4]) and biased sampling models (VARDI [6] ,[7]). We need to know more generally when these methods work as well as when they fail.

Dzhaparidze's generalization of Andersen and Gill's treatment of the time dependent covariate version of the Cox model is very far reaching. However, as in the i.i.d. Pareto model, as soon as the conditional hazard rates are no longer multiples of each other, we have an entirely different ball game. Even if the Pareto model is analyzable in the i.i.d. case in the way I suggested, analysis of its extensions to the general counting process framework remains a major challenge.

ADDITIONAL REFERENCES

1. P. K. ANDERSEN AND R. D. GILL (1982), Cox's regression model for counting processes: a large sample study, *Ann. Statist. 10*, 1100-1120.
2. K. R. BAILEY (1984), Asymptotic equivalence between the Cox estimator and the general ML estimators of regression and survival parameters in the Cox model, *Ann. Statist. 12*, 730-737.
3. R. A. BOYLES, A. W. MARSHALL AND F. PROSCHAN (1985), Inconsistency of the maximum likelihood estimator of a distribution having increasing failure rate average, *Ann. Statist. 13*, 413-417.
4. S. JOHANSEN (1983), An extension of Cox's regression model, *Int. Statist. Rev. 51*, 165-174.
5. A. A. TSIATIS (1981), A large sample study of Cox's regression model, *Ann. Statist. 9*, 93-108.
6. Y. VARDI (1982), Nonparametric estimation in the presence of length bias, *Ann. Statist. 10*, 616-620.
7. Y. VARDI (1985), Empirical Distributions in selection bias models, *Ann. Statist. 13*, 178-203.

Open discussion of papers on semiparametric models.

K. TAKEUCHI (JAPAN)
This question concerns the paper of *K. Dzhaparidze*. In order to get stronger results than the usual asymptotical normality of estimators, such as the convergence of moments, it seems that Condition IV, that is a kind of Lipschitz condition, is necessary, but I wonder whether it may put some restrictions on special cases such as Cox's model?

S.W. GREENHOUSE (USA)
I have two minor questions for *D. Clayton*. In section 3, you use a recursive procedure to find the maximum likelihood estimates of α and β. Do you have to check whether these do indeed maximize the likelihood jointly over the (α, β) space? Secondly, I believe α must have a certain sign a priori (either negative or positive). Is it possible to have a data set which yields an MLE of α which is of the wrong sign? If so, would not the semi-parametric procedure used subsequently lead you astray for that data set?

P. SOLOMON (UK)
This question to *D. Clayton* concerns the misspecification of regression models in the analysis of survival data. The proportional hazards and accelerated life families of regression models are widely used in the analysis of survival data. If interest lies in the qualitative effect on failure of various explanatory variables it is natural to ask how critical is the choice of model family in assessing the relative importance of the explanatory variables. It is known that if proportional hazards is assumed for analysis when the underlying distribution of failure time is accelerated life, then for small effects, the limit of the maximum likelihood estimate of the regression parameter under accelerated life is proportional to the true parameter. That is, the relative importance of the explanatory variables is preserved. Have you found this for your model?

R. HOGG (USA)
One possibly important application of *Clayton*'s methods is in casualty insurance, replacing survival times by amount of loss, given a loss has occurred. In terms of 'times', most of the loss distributions have decreasing failure rates and thus long tails on the right. In addition, these data are collected in groups; thus we have the interesting problem of not only having a group at the right end due to censoring, but we have a type of censoring throughout the data set.

D.R. Cox (UK)
Here are three questions. Firstly, (to *K. Dzhaparidze*) can Edgeworth expansions or saddlepoint expansions be obtained for the estimates under some condition? Secondly, (to *D. Clayton*) do you ever encounter divergent hazards in practice? Can time-dependent covariates be accommodated? Care is needed in interpreting the inflation of variance of β associated with not knowing α; cf. the recent discussion about estimating transformations in a more standard context. Orthogonality of α and β will help, also in the numerical analysis. Finally, (to *J.A. Wellner*) this classification of models is very helpful. Is it in general true that the asymptotically optimal semiparametric estimator is always asymptotically optimal for some parametric model (cf. EFRON [2] and OAKES [5], [6] on proportional hazards) and can such models be exhibited?

B. EPSTEIN (ISRAEL)
How do *Clayton*'s methods work when samples are small and there is substantial truncation?

J. GASTWIRTH (USA)
(to *D. Clayton*) Your paper contains a goodness of fit test of the proportional hazards model, namely is $\gamma = 0$ or not. Have you compared the power of your test with other procedures, such as the ones described in the recent report by GILL AND SCHUMACHER ([3]

E. SLUD (USA)
I would like to take exception here to the way in which Professors Clayton and Cuzick dismiss the Cox partial likelihood as a useful vehicle for the analysis of their semiparametric model. Unlike the Partial Likelihood in Cox's (1972, 1975) analysis of the proportional-hazards regression model, the analogous Partial Likelihood for the Clayton-Cuzick model will contain both the β and Λ_0 'parameters', but there is no reason to expect that the resulting expression cannot be jointly maximized with respect both to β and the nuisance hazard. (On the other hand, it is easy to check that only $\gamma \cdot \Lambda_0$, and not γ and Λ_0 separately, are identifiable from the partial likelihood.)

One motivation for taking a partial-likelihood approach, here as in Cox's (1972) regression model, is that if the distributions of the independent right-censoring variables are allowed to involve the parameters β and Λ_0 in some unspecified way, then the full or marginal likelihoods cannot even be written down until specific models for the censoring distributions are chosen. I believe that this justification for partial-likelihood inference has ordinarily not been brought forward in discussions of proportional-hazards regression because of the much more conspicuous computational advantages of having the nuisance hazard cancel out of the partial likelihood.

ANSWERS BY D. CLAYTON
(to *S.W. Greenhouse*) We have been led to this problem by an interest in epidemiology and in animal carcinogenesis studies. Here a Weibull model often fits rather well in practice. This suggests that the power law $\Lambda_0(t;\alpha) = t^\alpha$, $\alpha>0$, $\Lambda_0(t;0) = \log(1+t)$ might provide a useful generalisation of the basic parametric Pareto model. Negative values of α would correspond to analysis of reciprocals of survival times and this would transform right censoring into left censoring, but this is easily accomodated. Exclusion of negative α is not likely to be serious in practice since most of its effect would be on the sign of the β coefficient, and by not allowing negative α less confusion will arise. Regarding the point about convergence of the algorithm to the true MLE, when $\Lambda_0(t) = t^\alpha$ the likelihood is log-concave in α for all β and in β for all α (γ fixed), so that likelihood is increased at each step. However the likelihood is not generally log-concave in (α,β) jointly so saddle points may occur, and the only protection we know for this is to start the algorithm at different places. Monotone likelihoods are also possible which could lead to $\hat{\alpha} = 0$, $\hat{\alpha} = +\infty$ or some $|\beta_i| = +\infty$.

(to *P. Solomon*) If the accelerated failure model holds true, $\Lambda_0(t)$ belongs to the power law family

$\Lambda_0(t) = t^\alpha$. Clearly a non-parametric estimate of $\Lambda_0(t)$ is capable of adapting to this form. The error distribution in our model has a shape parameter, γ, which should allow it to adapt more freely to alternative error distributions than the proportional hazards model. Thus, one would conjecture that, if we are free to choose the most appropriate γ, then the estimates of β should be even more closely proportional to the equivalent parametric estimates than are the coefficients of the Cox model.

(to *D.R. Cox*) Our prime interest in this model arose out of a need to incorporate random effects in conventional proportional hazards regression analysis. A penalty of such random effects is, marginally, to destroy proportional hazards. That is, except in the case of the positive stable mixing distributions in which proportionality can be preserved, see HOUGAARD [4]. This paper exploits this side effect of our earlier work. However, there seems to be no way to allow for divergent hazards in this way. Of course, there are distributions of the error ϵ, which allow divergent hazards but the basic parametric kernel will no longer be expressable as a mixture of exponentials. Whether the regularity conditions which are necessary to support our approximations carry over into these cases remains to be seen. As regards time-dependent covariates, clearly this is an important extension which we should wish to be able to deal with. Unfortunately our present approach in terms of marginal likelihood seems rather limiting here, although formally the procedure can still be carried out. We agree that it is important to assess the loss of information associated with estimating α or more generally $\Lambda_0(t)$. When β is not near zero some loss will occur in general. It is of interest to note that when γ is also estimated the estimator for γ is independent of that for β when $\gamma = 0$, regardless of the value of β.

(to *P.J. Bickel*) We have found these suggestions most helpful and interesting and hope that they will lead to a more rigorous justification of our methods. However we suspect that a deep analytical problem lies at the base of this theory, and that this problem will be the same (or very similar) regardless of the basic framework used to set up the model.

(to *E. Slud*) Even in the proportional hazards case, Cox's partial likelihood is difficult to defend on theoretical grounds, although it certainly works well enough. A more satisfactory framework is the point process set up of Aalen, Gill and co-workers, see e.g. ANDERSEN et al [1]. This has the ability to accomodate very general censoring mechanisms, and hopefully someday will be useful for putting these more general models on a firm theoretical groundwork.

(to *J. Gastwirth*) GILL AND SCHUMACHER [3] have developed a class of tests for proportional hazards which consist of comparing different weighted averages of the relative risk function. In particular they consider weights corresponding to the 1-parameter family of tests studied by HARRINGTON AND FLEMING which include the Peto-Prentice generalization of the Wilcoxon test as a special case ($\gamma = 1$). The Pareto family of frailty models we considered yields the Harrington-Fleming test as score tests (weight functions) for the regression parameter when the (frailty) parameter is known. Also our test of $\gamma = 0$ is the locally most powerful test against this family of alternatives. It is not clear that the Gill-Schumacher tests are efficient for any model. However, we haven't investigated the behaviour of any of these tests for small samples.

(to *B. Epstein*) Our numerical work is limited, but the table in our paper indicates appreciable bias in the estimate when the sample size is twenty, but not when it is fifty. We suspect this is also true for the MLE in the parametric model. Our limited experiments with different types of censoring suggest that its main effect is to reduce the effective sample size to the number of uncensored observations, but when this exceeds fifty the bias will again be small.

ANSWERS BY K. DZHAPARIDZE

(to *K. Takeuchi*) A large sample study of the special Cox model usually requires asymptotic stability of the involved predictable processes, in the sense of ANDERSEN AND GILL (1982), for instance. Concerning this special model no conditions additional to those of the latter paper are required.

(to *D.R. Cox*) Results concerning some kind of expansions for the present type of problem are not known to me. This seems to be a very hard open research topic!

(to *P.Bickel*) In deriving some kind of asymptotic efficiency results the multiplicativity of the model in Aalen's sense (or 'asymptotic multiplicativity') indeed plays an important role.

ANSWERS BY J.A. WELLNER

(to *D.R. Cox*) At any particular point in the (infinite-dimensional) parameter space there is a class of locally equivalent hardest parametric submodels, for which the optimal semiparametric estimator remains optimal; the theory gives a recipe for exhibiting such models. So locally the answer is *yes*; but globally the question is not so meaningful since we are stuck with an infinite dimensional parametrization to fix points in the model.

ADDITIONAL REFERENCES

1. P. K. ANDERSEN, Ø. BORGAN, R. GILL AND N. KEILDING (1982), Linear nonparametric tests for comparison of counting processes, with applications to censored survival data, *International Statistical Review 49* , 219-258.
2. B. EFRON (1975), The efficiency of Cox's likelihood function for censored data, *J. Amer. Statist. Assoc. 72* , 557-565.
3. R. D. GILL AND M. SCHUMACHER (1985), *A simple test of the proportional Hazards Assumption* , C.W.I. Report MS-R8504, Amsterdam (to appear in Biometrika).
4. P. HOUGAARD (1984), Life table methods for heterogeneous populations: distributions describing the heterogeneity, *Biometrika 71* , 75-83.
5. D. OAKES (1977), The asymptotic information in censored survival data, *Biometrika 64* , 441-448. OAKES (1981), Survival Times: Aspects of Partial Likelihood, *International Statistical Review 49* , 235-264.

Efficient testing in a class of transformation models.

P.J. Bickel
Department of Mathematical Statistics
University of California at Berkeley[1]
Berkeley Ca. 94720
U.S.A.

1. Introduction.

Transformation models of the following type have been discussed among others by Cox [6], Clayton and Cuzick [4] and Doksum [8]. We observe (Z_i, Y_i) with $Y_i \in J_1$ an open subinterval of $\mathbb{R}$, which are a sample from a population characterized as follows. There exists an unknown transformation τ from J_0 an open subinterval of $\mathbb{R}$ onto J_1 with $\tau' > 0$ such that $Y = \tau(T)$ where (Z,T) follow a parametric model. The intervals J_i here may be proper or halfrays or $\mathbb{R}$ itself. Colloquially, if Y is expressed in the proper unknown scale, i.e. as T, then the joint behaviour of (Z,T) has some nice parametric form. The case considered by previous authors is

$$\log T = \theta^T Z + \epsilon$$

where ϵ is independent of Z. The distributions of ϵ considered so far include:

Cox [6]: e^{ϵ} has an exponential distribution.

Clayton and Cuzick [4]: e^{ϵ} has a Pareto distribution with density

$$f(t) = (1 + tc)^{-(\frac{1}{c}+1)}, \quad t > 0, \quad c \geq 0 \tag{1.1}$$

where $c = 0$ is the Cox model. An important special case of (1.1) considered by Bennett [2] is the log logistic model, $c = 1$ which has the attractive proportional odds property.

Doksum [8]: In generalization of the Box-Cox model, ϵ has a Gaussian distribution.

It seems reasonable in these models to base inference about the parameters of the underlying parametric model such as θ, c above on the maximal invariant of the group of transformations generating this semiparametric model, $\{(z,t) \to (z, \tau(t))\}$. This maximal invariant is just $M = (\underset{\sim}{Z}, \underset{\sim}{R})$ where $\underset{\sim}{Z} = (Z_1, \cdots, Z_N)$ and $\underset{\sim}{R} = (R_1, \cdots, R_N)$ is the vector of ranks of the Y_i. The likelihood of M or the conditional likelihood $L(\theta)$ of $\underset{\sim}{R}$ given $\underset{\sim}{Z} = \underset{\sim}{z}$ can in general only be expressed as an N dimensional integral. It can be evaluated explicitly for the Cox model. Clayton and Cuzick propose some ingenious approximations and Doksum proposes that both the value of L and its distribution be calculated approximately by Monte Carlo. So far, however, the asymptotic behaviour of these

1. Research supported by Office of Naval Research.

procedures is not well understood.

In this paper we specialize to $Z = 0,1$ as in BICKEL [3]. Moreover we suppose, as did Clayton and Cuzick, that the parameter θ governing the conditional density of $T = \tau(Y)$ given $Z = j$, denoted $f_j(\cdot,\theta)$ is real, and in particular that the distribution of ϵ is assumed known.

In this context, for a subclass of transformation models, we construct asymptotically efficient tests of $H: \theta = \theta_0$ vs $K: \theta > \theta_0$. The subclass includes the Pareto model for $c \geqslant 1$. The testing problem as such is not very interesting save in the case where θ_0 corresponds to independence of Y and Z which is already well understood. However, the solution of the testing problem is a first step in the solution of the estimation problem whose importance is clear. The tests we propose are based on 'quadratic rank statistics':

$$T_N = N^{-1} \sum_{i=1}^{N} a\left(\frac{R_i}{N}, Z_i\right) + N^{-2} \sum_{i,j=1}^{N} b\left(\frac{R_i}{N}, \frac{R_j}{N}, Z_i, Z_j\right). \tag{1.2}$$

We interpret efficiency in this context conditionally on $\underset{\sim}{Z}$, or equivalently the two sample sizes $\sum_{i=1}^{N} Z_i$ and $N - \sum_{i=1}^{N} Z_i$. We show,

i) If $\theta_N = \theta_0 + tN^{-1/2}$, $t \geqslant 0$,

$$\mathbf{L}_{\theta_N}\left(\frac{T_N}{\sigma_N} \mid \underset{\sim}{Z}\right) \to \mathbf{N}(at, 1) \text{ in probability for some } a > 0 \tag{1.3}$$

where σ_N is a sequence of normalizing constants.

ii) If S_N is any other sequence of statistics not necessarily depending on the ranks only such that

$$p\overline{\lim}_N \sup_\tau P_{(0,\tau)}[S_N \geqslant s \mid \underset{\sim}{Z}] = \alpha$$

then, for each τ, θ_N as above,

$$p\underline{\lim}_N P_{(\theta_N,\tau)}[S_N \geqslant s \mid \underset{\sim}{Z}] \leqslant 1 - \Phi(z_{1-\alpha} - at).$$

We use the subscripts θ and (θ,τ) to indicate the parameter values under which we compute. An important consequence of (i) and (ii) is the following. Let $\Lambda_N(t) = L(\theta_N)/L(\theta_0)$ be the conditional rank likelihood ratio statistic for $H: \theta = \theta_0$ vs $K: \theta = \theta_N$ given $\underset{\sim}{Z}$. Then $L(\theta_N)/L(\theta_0)$ is also efficient. That is an asymptotically size α test based on $\Lambda_N(t)$ has power given by (1.3) and in fact tests based on $\Lambda_N(t)$ for different t are asymptotically equivalent to each other and our quadratic rank test.

We begin by heuristically deriving what turn out to be appropriate a and b. In section 2 we discuss existence and computation of T_N, in section 3 asymptotic normality for T_N, and in section 4, efficiency. Extensions to estimation, the Pareto model for $0 \leqslant c < 1$ and the normal model as well as to censored data are discussed in section 5. We also present some preliminary Monte Carlo results in this section.

We first calculate formally the locally most powerful rank test statistic for $H: \theta = \theta_0$ vs $K: \theta > \theta_0$. By Hoeffding's formula,

$$P_\theta[\underset{\sim}{R} = \underset{\sim}{r} \mid \underset{\sim}{Z} = \underset{\sim}{z}] = E_{\theta_0}\left\{\prod_{i=1}^{N}\left(\frac{f_0(T_i,\theta)}{f_0(T_i,\theta_0)}\right)^{1-z_i}\left(\frac{f_1(T_i,\theta)}{f_1(T_i,\theta_0)}\right)^{z_i} I(\underset{\sim}{R} = \underset{\sim}{r})\right\}$$

so that the locally most powerful test statistic

$$N^{-\frac{1}{2}} \frac{\partial}{\partial\theta} \log P_\theta[\underset{\sim}{R} = \underset{\sim}{r} \mid \underset{\sim}{Z} = \underset{\sim}{z}]$$

is just $N^{\frac{1}{2}} S_N$ where

$$S_N = N^{-1} \sum_{i=1}^{N} Z_{i0} E_{\theta_0}\{c_0(T_i) \mid \underset{\sim}{Z}, \underset{\sim}{R}\} + Z_{i1} E_{\theta_0}\{c_1(T_i) \mid \underset{\sim}{Z}, \underset{\sim}{R}\}$$

where $c_j(t) = \frac{\partial \log}{\partial \theta} f_j(t, \theta_0)$ and $Z_{ij} = I(Z_i = j)$. Equivalently, if $\underset{\sim}{D} = (D_1, \cdots, D_N)$ are the antiranks defined by $T_{(j)} = T_{D_j}$ where $T_{(1)} < \cdots < T_{(N)}$ are the order statistics of the sample, then

$$S_N = N^{-1} \sum_{j=0}^{1} \sum_{i=1}^{N} \{Z_{D_i j} E_{\theta_0}(c_j(T_{(i)}) \mid \underset{\sim}{Z}, \underset{\sim}{D})\}. \tag{1.4}$$

To get an approximation to the scores in (1.4) we write $f_j(\cdot, \theta_0)$ as $f_j(\cdot)$, and define

$$n = \sum_{i=1}^{N} Z_i, \quad m = N - n, \quad \hat{\pi}_0 = \frac{m}{N} = 1 - \hat{\pi}_1.$$

We treat $\hat{\pi}_j$ as deterministic constants in the sequel. Let

$$h(\cdot) = \hat{\pi}_0 f_0(\cdot) + \hat{\pi}_1 f_1(\cdot)$$

with H the corresponding distribution function. Note that h and H depend on N and are random only through the $\hat{\pi}_j$.

Finally let, for $0 < t < 1$,

$$\lambda_j(t) = c_j(H^{-1}(t)), \tag{1.5}$$

$$g_j(t) = f_j(H^{-1}(t)) / h(H^{-1}(t))$$

the density of $H(T_1)$ given $Z_1 = j$, and

$$\gamma_j(t) = -\frac{g_j'}{g_j}(t). \tag{1.6}$$

We can rewrite (1.4) as

$$S_N = S_{N1} + S_{N0}$$

where

$$S_{Nj} = N^{-1} \sum_{i=1}^{N} Z_{D_i j} E(\lambda_j(U_{(i)}) \mid \underset{\sim}{Z}, \underset{\sim}{D})$$

where (Z_i, U_i) are i.i.d with U_1 given $Z_1 = j$ having density g_j and the marginal density of U_1 is uniform,

$$\hat{\pi}_0 g_0 + \hat{\pi}_1 g_1 = 1. \tag{1.7}$$

The next step is to note that $U_{(i)} \approx \frac{i}{N}$ so that

$$S_{Nj} \approx N^{-1} \sum_{i=1}^{N} \{Z_{D_i j} (\lambda_j(\tfrac{i}{N}) + \lambda_j'(\tfrac{i}{N}) E[(U_{(i)} - \tfrac{i}{N}) \mid \underset{\sim}{Z}, \underset{\sim}{D}]\} \tag{1.8}$$

plus terms we expect to be of order $\mathcal{O}(N^{-1})$.

The first term of the approximation is a linear rank statistic. For the second we use a heuristic argument of Clayton and Cuzick who argue that if

$$\bar{y}_i = E(U_{(i)} \mid \underset{\sim}{Z}, \underset{\sim}{D})$$

then $\bar{y}_i$ satisfies approximately the recurrence relation,

$$-\{(\bar{y}_{i+1} - \bar{y}_i)^{-1} - (\bar{y}_i - \bar{y}_{i-1})^{-1}\} = (1 - Z_{D_i}) \gamma_0(\bar{y}_i) + Z_{D_i} \gamma_1(\bar{y}_i). \tag{1.9}$$

Let

$$\hat{G}_j(t) = (N \hat{\pi}_j)^{-1} \sum_{i=1}^{N} I(U_i \leqslant t) Z_{ij}$$

be the empirical d.f.s of the two subsamples of U_i from g_0, g_1, and let

$$\hat{R}(t) = \pi_N \hat{G}_0 + (1-\pi_N)\hat{G}_1 \tag{1.10}$$

be the empirical d.f. of the sample $U_1, \cdots, U_N$. Define,

$$\hat{Q}_j(t) = \hat{G}_j \hat{R}^{-1}(t+0), \quad 0 < t \leqslant 1 \tag{1.11}$$
$$\hat{Q}_j(0-) = 0$$

where for any d.f. F, $F^{-1}(t) = \inf\{s : F(s) \geqslant t\}$. $\hat{Q}_0$ is a distribution function with jumps of size m^{-1} at $\frac{j-1}{N}$ such that $Z_{D_j} = 0$ while $\hat{Q}_1$ jumps $(N-m)^{-1}$ at $\frac{j-1}{N}$ with $Z_{D_j} = 1$. Evidently $\bar{y}_i$ is a function of $\frac{i}{N}$, $\underset{\sim}{Z}$, $\hat{Q}_0$, $\hat{Q}_1$ only. Interpolate smoothly in some way between $\frac{i-1}{N}$ and $\frac{i}{N}$, $1 \leqslant i \leqslant N$ to obtain a function v on (0,1) such that,

$$\bar{y}_i = v\left(\frac{i}{N}\right).$$

Any solution of (1.9) must satisfy, for some c,d

$$\bar{y}_i = d + \sum_{j=1}^{i} (c + \sum_{k>j} Z_{D_k 0}\gamma_0(\bar{y}_k) + Z_{D_k 1}\gamma_1(\bar{y}_k))^{-1}$$

or for $u = \frac{i}{N} \approx \frac{i-1}{N}$

$$v(u) \approx d + \int_0^u \left(\frac{c}{N} + \int_t^1 \gamma_0(v(s))\hat{\pi}_0 d\hat{Q}_0(s) + \gamma_1(v(s))\hat{\pi}_1 d\hat{Q}_1(s)\right)^{-1} dt. \tag{1.12}$$

This is essentially the integral equation of BICKEL [3], save that we make the transformation $H(\cdot)$ and apply (1.8). Unfortunately, the hopes for analytic approximation of solutions to (1.12) expressed in BICKEL [3] have so far not been realized. However, suppose we (still formally) extend the definition of (1.12) to functions $v(\cdot, Q, Q')$ by replacing $\hat{Q}_0$ $\hat{Q}_1$ by arbitrary Q, Q' such that,

$$\hat{\pi}_0 Q(t) + \hat{\pi}_1 Q'(t) = t, \quad \text{for } t = 0, \frac{1}{N}, \cdots, 1$$

with c,d depending on Q, Q'. Then, if $Q = G_0$, $Q' = G_1$, $\frac{c}{N} = 1$ and $d = 0$, $v(u) = u$ formally satisfies the extension of (1.12) since by (1.7)

$$\gamma_0 \hat{\pi}_0 g_0 + \gamma_1 \hat{\pi}_1 g_1 = 0.$$

Therefore, writing $v(u) = v(u, \hat{Q}_0, \hat{Q}_1)$

$$\tilde{\Delta}(u) = v(u) - u = \int_0^u \{[c + \int_t^1 (\gamma_0(v(s))\hat{\pi}_0 d\hat{Q}_0(s) + \gamma_1(v(s))\hat{\pi}_1 d\hat{Q}_1(s))]^{-1} dt$$
$$- [1 - \int_t^1 (\gamma_0(s)\hat{\pi}_0 dG_0(s) + \gamma_1(s)\hat{\pi}_1 dG_1(s))]^{-1}\} dt\}.$$

The constants $c(\hat{Q}_0, \hat{Q}_1)$, $d(\hat{Q}_0, \hat{Q}_1)$ are formally determined by smooth fit at the boundaries,

$$\tilde{\Delta}(0) = \tilde{\Delta}(1) = 0. \tag{1.13}$$

Let,

$$\alpha(s) = \sum_{j=0}^{1} \gamma_j'(s)\hat{\pi}_j g_j(s). \tag{1.14}$$

Then,

$$\tilde{\Delta}(u) \approx -\int_0^u \{\int_t^1 [\hat{\pi}_0(\gamma_0(v(s))d\hat{Q}_0(s) - \gamma_0(s)dG_0(s)) + \hat{\pi}_1(\gamma_1(v(s))d\hat{Q}_1(s) \tag{1.15}$$

$$-\gamma_1(s)dG_1(s)]\}dt + (c(\hat{Q}_0,\hat{Q}_1) - 1)u + d(\hat{Q}_0,\hat{Q}_1)$$

$$\approx -\int_0^u \{\int_t^1 \alpha(s)\tilde{\Delta}(s)ds + \int_t^1 \gamma_0(s)\hat{\pi}_0 d(\hat{Q}_0(s) - G_0(s)) + \gamma_1(s)\hat{\pi}_1 d(\hat{Q}_1(s) - G_1(s))]dt$$

$$+(c(\hat{Q}_0,\hat{Q}_1) - 1)u + d(\hat{Q}_0,\hat{Q}_1)$$

$$\approx -\int_0^u \int_t^1 \alpha(s)\tilde{\Delta}(s)ds + \int_t^1 \sum_{j=0}^1 \gamma_j(s)\hat{\pi}_j d(\hat{Q}_j - G_j)(s)$$

$$+ u\int_0^1 (\int_v^1 \alpha(s)\tilde{\Delta}(s)ds + \int_v^1 \sum_{j=0}^1 \gamma_j(s)\hat{\pi}_j d(\hat{Q}_j - G_j)(s)).$$

After some algebra, this reduces to,

$$\tilde{\Delta}(u) \approx -\int_0^1 K(s,u)\alpha(s)\tilde{\Delta}(s)ds - \int_0^1 K(s,u)\sum_{j=0}^1 \gamma_j(s)\hat{\pi}_j d\hat{Q}_j(s) \tag{1.16}$$

where

$$K(s,u) = s\Lambda u - su.$$

Continuing to ignore existence and unicity questions we define $\Delta(u)$ as the solution of the linear integral equation obtained from the approximate equation (1.16). We introduce a Green's function solving

$$\Delta(u,v) + \int_0^1 K(s,u)\alpha(s)\Delta(s,v)ds = K(u,v). \tag{1.17}$$

Then $\Delta(t)$ is given by

$$\Delta(t) = -\int_0^1 \Delta(t,u)\sum_{j=0}^1 \hat{\pi}_j\gamma_j(u)d\hat{Q}_j(u). \tag{1.18}$$

We now define

$$T_N = \int_0^1 \sum_{j=0}^1 (\lambda_j(t) + \lambda_j'(t)\Delta(t))\hat{\pi}_j d\hat{Q}_j(t) \tag{1.19}$$

$$= \sum_{j=0}^1 \int_0^1 \lambda_j(t)\hat{\pi}_j d\hat{Q}_j(t) - \sum_{j=0}^1 \sum_{k=0}^1 \int_0^1 \int_0^1 \Delta(t,u)\lambda_j'(t)\gamma_k(u)\hat{\pi}_j\hat{\pi}_k d\hat{Q}_j(t)d\hat{Q}_k(u)\,,$$

which is of the form (1.2) with

$$a(\frac{i}{N},j) = \lambda_j(\frac{i-1}{N})$$

$$b(\frac{i}{N},\frac{i'}{N},j,k) = \lambda_j'(\frac{i-1}{N})\gamma_k(\frac{i-1}{N})\Delta(\frac{i-1}{N},\frac{i'-1}{N}).$$

If

$$\int_0^1 |\gamma_j'(t)|dt < \infty,\ j = 0,1\,, \tag{1.20}$$

we shall see in Lemma 2.2 that Δ defined by (1.17) exists and is unique. We sketch in section 3, the asymptotic theory of T_N given by (1.19) under (1.20) and

$$\int_0^1 |\lambda_j''(t)|\,dt < \infty \quad j = 0,1 \tag{1.21}$$

Conditions (1.20) and (1.21) are satisfied for the Pareto family if $c \geq 1$. They fail for $c < 1$ and the normal and exponential families.

2. Existence and computation of T_N.

We establish existence and unicity and some properties of $\Delta(t,u)$ given by (1.17).

Lemma 2.1. *Suppose f_0, f_1 have common support $S = \{f_0 > 0\}$ and are twice continuously differentiable on S. Then, γ_0, γ_1 are continuously differentiable on $(0,1)$, α is continuous and,*

$$\alpha = \sum_{j=0}^{1} \hat{\pi}_j g_j \gamma_j^2 \tag{2.1}$$

so that,

$$\alpha(s) \geq 0 \quad \text{for } 0 < s < 1, \tag{2.2}$$

with equality iff $\dfrac{f_0'}{f_0}(H^{-1}(s)) = \dfrac{f_1'}{f_1}(H^{-1}(s))$.

Proof. By (1.7)

$$\sum_{j=0}^{1} \pi_j \gamma_j g_j = 0. \tag{2.3}$$

Differentiating,

$$\alpha = \sum_{j=0}^{1} \pi_j \gamma_j' g_j = -\sum_{j=0}^{1} \pi_j g_j' \gamma_j = \sum_{j=0}^{1} \pi_j \gamma_j^2 g_j. \tag{2.4}$$

Further,

$$g_j = \frac{f_j}{h}(H^{-1}) \tag{2.5}$$

so that

$$\gamma_j = \frac{1}{h}\left(\frac{f_j'}{f_j} - \frac{h'}{h}\right)(H^{-1}) \tag{2.6}$$

and $\gamma_j = 0$ for all j iff all the $\dfrac{f_j'}{f_j}(H^{-1}(s))$ are equal.

□

Lemma 2.2. *Suppose that the conditions of lemma 2.1 hold. Then, Δ defined by (1.17) exists and is unique. Moreover*

$$\int_0^1 \Delta^2(u,v)\alpha(u)\,du \leq \int_0^1 \alpha(u)\,du \tag{2.7}$$

$$|\Delta(u,v)| \leq \int_0^1 \alpha(u)\,du + 1 \tag{2.8}$$

Proof. By (1.20), γ_j are bounded on $[0,1]$ and by (2.1) so is α. As in Tricomi [12] p. 3 let,

$$\psi(u,v) = \sqrt{\alpha(u)}\,\Delta(u,v).$$

Then, ψ satisfies

$$L(\psi(\cdot,v))(u) = \sqrt{\alpha(u)}K(u,v)$$

where L is the operator on $L_2(0,1)$ given by,

$$L(\psi) = I + T$$

where I is the identity and

$$K(\psi)(t) = \int_0^1 \sqrt{\alpha(s)}K(s,t)\sqrt{\alpha(t)}\psi(s)ds.$$

Since $\alpha \in L_2(0,1)$ the operator K is bounded, self adjoint and nonnegative definite since the kernel K, the covariance kernel of the Brownian bridge, is. Hence, L is 1-1 and onto. Moreover, all eigenvalues of L are ≥ 1 so that $\|L^{-1}\| \leq 1$. Existence, uniqueness and (2.7) follows. Further from (1.16)

$$|\Delta(t,u)| \leq \int_0^1 |\Delta(s,u)|\,\alpha(s)K(s,t)ds + K(t\Lambda u) \leq \left(\int_0^1 \Delta^2(s,u)\alpha(s)ds\right)^{\frac{1}{2}} \left(\int_0^1 \alpha(s)ds\right)^{\frac{1}{2}} + 1.$$

□

LEMMA 2.3. *(a). Under the conditions of lemma 2.1,* $\Delta(\cdot,u)$ *is continuously differentiable on* $[0,u),(u,1]$ *and* $\frac{\partial\Delta}{\partial t}(t,u)$ *has a jump discontinuity of -1 at* $t = u$. *Moreover,* $\Delta(\cdot,u)$ *is the unique solution of the Sturm-Liouville equation,*

$$y''(t) - \alpha(t)y(t) = 0 \tag{2.9}$$

everywhere except at $t = u$, *which satisfies the boundary conditions,*

$$a)\ y(0) = 0 \tag{2.10}$$

$$b)\ y(1) = 0$$

(b). Suppose y_1, y_2 *are fundamental solutions of (2.9) satisfying* $y_1(0) = 0$, $y_1'(0) = 1$, $y_2(1) = 0$, $y_2'(1) = -1$ *say. Then* y_1, y_2 *are linearly independent and* $\Delta(t,u)$ *is given by,*

$$\Delta(t,u) = \frac{y_1(t)y_2(u)}{D(u)} \quad 0 \leq t \leq u \tag{2.11}$$

$$= \frac{y_2(t)y_1(u)}{D(u)}, \quad u \leq t \leq 1$$

where $D(u) \equiv y_2(0)$ *is, the Wronskian of (2.9).*

PROOF. From (1.17), $\Delta(\cdot,u)$ is absolutely continuous and satisfies,

$$\frac{\partial\Delta}{\partial t}(t,u) + \int_t^1 \alpha(s)\Delta(s,u)ds = \int_0^1 s\alpha(s)\Delta(s,u)ds + I(t \leq u) - u. \tag{2.12}$$

Differentiating again for $t \neq u$ we obtain (2.9). The boundary conditions follow from (1.17).

If $y(\cdot,u)$ is a 'solution' of (2.9) as above it can be integrated twice to obtain (1.17). Finally, it is easy to verify that (2.11) is a 'solution' of (2.8) satisfying (2.9) and (2.10) with the required jump discontinuity. The background for these calculations may be found in HILLE [11], Theorem 8.2.1 and lemma 8.5.1.

□

3. ASYMPTOTIC LINEARITY AND NORMALITY.

We give in this section linear approximations to two sample quadratic rank statistics which are uniform over families of statistics as well as families of distributions. The generality is greater than we need for this paper but will be needed in the sequel treating models in which α has a singularity at 0 or 1. We essentially use the methods of CHERNOFF AND SAVAGE [5] and PYKE AND SHORACK [13].

Suppose $U_1, \cdots, U_m$ are i.i.d G_0, $U_{m+1}, \cdots, U_N$ are i.i.d G_1, $N = m + n$. Let $\hat{\pi}_0 = \frac{m}{N} = 1 - \hat{\pi}_1$ and suppose G_j have twice continuously differentiable densities g_j on (0,1) such that,

$$\hat{\pi}_0 g_0 + \hat{\pi}_1 g_1 = 1.$$

G_0, G_1 can vary with N.

Define γ_j as in (1.5), $\hat{Q}_j$, $\hat{R}$ as in (1.10), (1.11). For given a,c: $[0,1] \to R$, $k : [0,1] \to R^+$ let

$$S_N = \int_0^1 a(t) d\hat{Q}_j(t) \tag{3.1}$$

$$V_N = \int_0^1 \int_0^1 b(s,t) d\hat{Q}_j(u) d\hat{Q}_k(v) \tag{3.2}$$

where

$$b(s,t) = a(s)c(t)\Delta(s,t),$$

$$\Delta(t,u) + \int_0^1 K(s,t)k(s)\Delta(s,u)ds = K(t,u). \tag{3.3}$$

All the results of Section 2 apply to Δ given in (3.3) when we replace α by k since no property of α other than $\alpha \geqslant 0$ was used in that section. Let

$$\mathbf{P}_M = \{P \mid \|\gamma_j\|_1 \leqslant M, \ \|\log g_j\|_\infty \leqslant M, \ 0 < \epsilon \leqslant \hat{\pi}_0 \leqslant 1 - \epsilon\}.$$

$$\mathbf{A}_M = \{a \mid a' \text{absolutely continuous}, \ \|a'\|_1 + \|a\|_\infty \leqslant M\}$$

$$\mathbf{B}_M = \{k \mid \|k\|_1 \leqslant M^2\},$$

where $\|\cdot\|_p$ is the L_p norm on (0,1) or $(0,1)\times(0,1)$ as appropriate.

If $R_N(a,P)$ is a statistic based on $U_1, \cdots, U_N$ and P denotes the probability measure corresponding to (G_0, G_1) as well, and for any $\epsilon \in 0$, there exist $K(\epsilon) < \infty$, $N(\epsilon) < \infty$ independent of P, a, M such that,

$$P[|R_N(a,P)| \leqslant K(\epsilon)c(M)d(N)] \geqslant 1 - \epsilon, \text{ if } N \geqslant N(\epsilon)$$

we shall write,

$$R_N = \mathcal{O}(c(M)d(N)) \tag{3.4}$$

LEMMA 3.1 *If* $a \in \mathbf{A}$, $P \in \mathbf{P}$, *write*

$$S_N = \int_0^1 a(t) dG_j(t) + \int_0^1 a(t) d(\hat{G}_j - G_j)(t) + \int_0^1 B_j(t) d(\hat{R}(t) - t) + R_N(a,P),$$

where

$$B_j(t) = \int_t^1 a'(s) g_j(s) ds.$$

Then, for every $\delta > 0$,

$$R_N = \mathcal{O}(N^{-\frac{3}{4}+\delta} M).$$

PROOF. Since $a \in \mathbf{A}$, $P \in \mathbf{P}$, so that

$$S_N = \int_0^1 a(t)dG_j(t) - \int_0^1 (\hat{Q}_j(t) - G_j(t))a'(t)dt - a(0)\hat{Q}_j(0). \tag{3.5}$$

By strong approximation theorems to the quantile and empirical processes , see e.g. CSÖRGÖ [7], and a standard estimate on the modulus of continuity of the Brownian bridge,

$$\|(\hat{Q}_j(\cdot) - G_j(\cdot)) - (\hat{G}_j(\cdot) - G_j(\cdot)) + g_j(\cdot)(\hat{R}(\cdot) - \cdot)\|_\infty = \mathcal{O}(N^{-\frac{3}{4}+\delta} Me^M).$$

We use here that $\|g_j'\|_1 \leq Me^M$. Therefore, by (3.5)

$$S_N = \int_0^1 (\hat{G}_j(t) - G_j(t))a'(t)dt - \int_0^1 g_0(t)(\hat{R}(t) - t)a'(t)dt + O(N^{-\frac{3}{4}+\delta} Me^M)$$

since $|a(0)\hat{Q}_j(0)| \leq MN^{-1}$.

□

LEMMA 3.2 *Suppose* $a,c,d \in \mathbf{A}_M$, $k \in \mathbf{B}_M$, $P \in \mathbf{P}$. *Write*

$$V_N = \int_0^1 \int_0^1 b(s,t)dG_j(s)dG_k(t) + \int_0^1 A_1(s)d(\hat{G}_j - G_j)(s) \tag{3.6}$$

$$+ \int_0^1 A_2(t)d(\hat{G}_k(t) - G_k(t)) + \int_0^1 B(t)d(\hat{R}(t) - t) + R_N(b,P)$$

where

$$A_1(s) = a(s)\int_0^1 \Delta(s,t)dG_k(t) \tag{3.7}$$

$$A_2(t) = c(t)\int_0^1 \Delta(s,t)dG_j(s) \tag{3.8}$$

$$B(u) = \int_u^1 (A_1'(s)g_j(s) + A_2'(s)g_k(s))ds. \tag{3.9}$$

Then

$$R_N = O(N^{-\frac{3}{4}+\delta} Me^M)$$

PROOF. Write

$$V_N = \int_0^1 \int_0^1 b(s,t)dG_j(s)dG_k(t) + \int_0^1 A_1(s)d(\hat{Q}_j - G_j)(s) \tag{3.10}$$

$$+ \int_0^1 A_2(t)d(\hat{Q}_k - G_k)(t) + \int_0^1 \int_0^1 b(s,t)d(\hat{Q}_j - G_j)(s)d(\hat{Q}_k - G_k)(t),$$

and let

$$W_1(u) = \int_u^1 a(s)d(\hat{Q}_j - G_j)(s),$$

$$W_2(v) = \int_v^1 c(t)d(\hat{Q}_k - G_k)(t),$$

$$d(t) = \int_0^1 sk(s)\Delta(s,t)ds. \tag{3.11}$$

Then

$$\int_0^1\int_0^1 b(s,t)d(\hat{Q}_j - G_j)(s)d(\hat{Q}_k - G_k)(t) \tag{3.12}$$
$$= \int_0^1\int_0^1 \Delta(s,t)dW_1(s)dW_2(t)$$
$$= -\int_0^1\Big(\int_0^1 \frac{\partial\Delta}{\partial s}(s,t)W_1(s)ds\Big)dW_2(t)$$
$$= \int_0^1\int_0^1\Big\{\Big(\int_s^1\Delta(v,t)k(v)dv\Big) - I(s \leqslant t) - \big(d(t)-t\big)\Big\}W_1(s)dsdW_2(t)$$

by (2.12) with $\alpha = k$. The first term in (3.12) above is

$$= \int_0^1\int_0^1\Big(\int_0^v W_1(s)ds\Big)\Delta(v,t)k(v)dW_2(t)dv. \tag{3.13}$$

By lemma 2.3 since Δ is symmetric

$$\frac{\partial}{\partial t}\Delta(v,t) + \int_t^1\Delta(s,v)k(s)ds = I(t \leqslant v) + d(v) - v. \tag{3.14}$$

Substituting into (3.13) and then (3.12) we get

$$\int_0^1\int_0^1 b(s,t)d(\hat{Q}_j - G_j)(s)d(\hat{Q}_k - G_k)(t) \tag{3.15}$$
$$= \int_0^1\int_0^1\Big\{\Big(\int_0^v W_1(s)ds\Big)k(v)\Big[\int_0^u W_2(t)dt\Big)\Delta(u,v)k(u)$$
$$-(W_2(v) - W_2(0) + (d(v)-v)W_2(0)\Big]\Big\}dudv$$
$$-\int_0^1 W_1(s)W_2(s)ds + \Big(\int_0^1 W_1(s)ds\Big)\Big(\int_0^1 W_2(t)\big(d'(t)-1\big)dt\Big).$$

Argueing as in lemmas 3.1,

$$\|W_1\|_\infty = \mathcal{O}(N^{-\frac{1}{2}}(\|a'\|_1 + \|a\|_\infty)) \tag{3.16}$$
$$\|W_2\|_\infty = \mathcal{O}(N^{-\frac{1}{2}}(\|c'\|_1 + \|c\|_\infty)).$$

Moreover since

$$d(t) + \int_0^1 K(t,s)k(s)d(s)ds = \int_0^1 K(t,s)sk(s)ds,$$

arguing as in lemma 2.2

$$\|d\|_\infty \leqslant \Big(\int_0^1 k(s)d^2(s)ds\Big)^{\frac{1}{2}}\Big(\int_0^1 k(s)ds\Big)^{\frac{1}{2}} + \int_0^1 k(s)ds \leqslant 2\int_0^1 k(s)ds. \tag{3.17}$$

Also

$$d'(t) = -\int_t^1 K(t,s)k(s)d(s)ds + \int_t^1 sk(s)ds - \int_0^1 s^2k(s)ds$$

and

$$\|d'\|_\infty \leq 2\int_0^1 k(s)ds. \tag{3.18}$$

Combining (3.16) - (3.18) permits us to bound (3.15) by

$$\begin{aligned} &\mathcal{O}(N^{-1}M^2\Big[\int_0^1\int_0^1 k(u)k(v)\Delta(u,v)dudv \qquad (3.19)\\ &+ \Big(\int_0^1 k(v)dv\Big)\Big(1 + \int_0^1 k(v)dv\Big) + \Big(1 + \int_0^1 k(v)dv\Big)\Big]\Big)\\ &= \mathcal{O}(N^{-1}(1 + M^6)) \end{aligned}$$

since $k \in \mathbf{B}_M$.

Apply lemma 3.1 to the first three terms in (3.10) and (3.19) to the last to obtain the lemma. □

THEOREM 3.1 *Suppose T_N is given by (1.19) where $\alpha \in \mathbf{B}_M$ and γ_j, λ_j, $\lambda_j' \in \mathbf{A}_M$. Then*

$$T_N = \sum_{j=0}^1 \hat{\pi}_j \int_0^1 \lambda_j(t)dG_j(t) + N^{-1}\sum_{i=1}^N\sum_{j=0}^1 A_j(U_i)Z_{ij} + \mathcal{O}(N^{-\frac{3}{4}+\delta}Me^M) \tag{3.20}$$

where

$$A_j(u) = \lambda_j(u) - \int_0^1 \lambda_j(u)dG_j - \sum_{l=0}^1 (v_l'(u) - \gamma_j(u)v_l(u)) \tag{3.21}$$

and

$$v_j(u) = \hat{\pi}_j \int_0^1 \Delta(t,u)\lambda_j'(t)dG_j(t). \tag{3.22}$$

PROOF. Note that by assumption A_j are bounded. Moreover,

$$\begin{aligned} EA_j(U_1) &= -\sum_{l=0}^1 \Big(\int_0^1 v_l'(u)du + \int_0^1 (\hat{\pi}_0\gamma_0 g_0(u) + \hat{\pi}_1\gamma_1 g_1(u))v_l(u)du\\ &= \sum_{l=0}^1 (v_l(1) - v_l(0)) = 0. \end{aligned}$$

Apply lemmas 3.2 and 3.1 to the quadratic and linear parts of T_N respectively and obtain that T_N is linear in $\hat{\pi}_j(\hat{G}_j - G_j)$, $j = 0,1$, and $\hat{R}(\cdot) - \cdot)$ with remainder $\mathcal{O}(N^{-\frac{3}{4}+\delta}Me^M)$. We show that the linear part can be put in the form (3.20). By symmetry it suffices to argue this if one of the λ_j, λ_0 say, is identical to 0. Then, from lemmas 3.1 and 3.2, after some simplification,

$$\begin{aligned} T_N = \hat{\pi}_1\Big(&\int_0^1 \lambda_1(t)dG_1(t) + \int_0^1 \lambda_1(t)d(\hat{G}_1 - G_1)(t) \qquad (3.23)\\ &+ \int_0^1\Big(\int_s^1 \lambda'(t)dG_1(t)\Big)d(\hat{R}(s) - s)\end{aligned}$$

$$-\sum_{j=0}^{1}\int_0^1\gamma_j(u)\Big(\int_0^1\Delta(t,u)\lambda_1'(t)dG_1(t)\Big)\hat{\pi}_j d(\hat{G}_j - G_j)(u)$$

$$-\sum_{j=0}^{1}\int_0^1\int_s^1\frac{\partial}{\partial u}\Big\{\gamma_j(u)\int_0^1\Delta(t,u)\lambda_1'(t)dG_1(t)\Big\}\hat{\pi}_j dG_j(u)d(\hat{R}(s)-s))+\mathcal{O}(N^{-\frac{3}{4}+\delta}Me^{-M}).$$

Now, since

$$\frac{\partial}{\partial u}\gamma_j(u)\int_0^1\Delta(t,u)\lambda_1'(t)dt = \gamma_j'(u)\int_0^1\Delta(t,u)\lambda_1'(t)dt + \gamma_j(u)\int_0^1\frac{\partial\Delta}{\partial u}(t,u)\lambda_1'(t)dt$$

and $\sum_{j=0}^{1}\gamma_j(u)\hat{\pi}_j g_j(u) = 0$, the last term in (3.23) simplifies to

$$-\int_0^1\Big(\int_s^1\alpha(u)\Delta(t,u)du\lambda_1'(t)dG_1(t)\Big)d(\hat{R}(s)-s). \tag{3.24}$$

Since Δ is symmetric, (3.24) simplifies using (2.12) to

$$\int_0^1\Big(\frac{\partial}{\partial s}\Delta(s,t) - I(s\leq t) - d(t) + t\Big)\lambda_1'(t)G_1(t)d(\hat{R}(s)-s)$$

where d is given by (3.11) with $k = \alpha$. We finally obtain

$$\int_0^1\Big(\frac{\partial}{\partial s}\int_0^1\Delta(s,t)\lambda_1'(t)dG_1(t) - \int_s^1\lambda_1'(t)dG_1(t)\Big)d(\hat{R}(s)-s).$$

After substituting in (3.23) we arrive at (3.20).

□

4. Efficiency.

We are throughout conditioning on Z, i.e. treating the situation where we have two samples of known size m and N-m, from $F_1(\tau^{-1}(\cdot),\theta)$ and $F_0(\tau^{-1}(\cdot),\theta)$ respectively. Without loss of generality suppose that under the basic model, conditionally on $Z = j$, T has the distributions on (0,1) given by

$$g_j(t,\theta) = \frac{f_j(H^{-1}(t),\theta)}{h(H^{-1}(t))} \tag{4.1}$$

and suppose that

A: All of these distributions have common support (0,1) and

$$\lambda_j(t,\theta) = \frac{\partial}{\partial\theta}\log g_j(t,\theta) \tag{4.2}$$

$$\gamma_j(t,\theta) = -\frac{g_j'}{g_j}(t,\theta) \tag{4.3}$$

are well defined. Moreover, we require $\lambda_j(\cdot,\theta)$, $\lambda_j'(\cdot,\theta)$, $\gamma_j(\cdot,\theta)$ to belong to $\mathbf{A}_M$ for some M, θ close enought to θ_0.

Since

$$\lambda_j(t,\theta_0) = \lambda_j(t),\quad \gamma_j(t,\theta_0) = \gamma_j(t),$$

we see by lemma 2.1, that, under A, $\alpha\in\mathbf{B}_M$ and all conditions of theorem 3.1 are satisfied.

These conditions are easily seen to be satisfied for the Pareto model if $c\geq 1$. In that case, in fact, $\gamma'(\cdot,\theta)$, $\lambda''(\cdot,\theta)$ are continuous and a fortiori bounded on [0,1].

Under A, if v_j is defined by (3.22),

$$\|v_j'\|_\infty = \mathcal{O}(M^3). \tag{4.4}$$

Let,

$$q(t,\theta) = t - (\theta - \theta_0)\sum_{j=0}^{1} v_j(t), \quad 0 \leq t \leq 1, \tag{4.5}$$

if $|\theta - \theta_0| < (\sum_{j=0}^{1} \|v_j'\|_\infty)^{-1}$. Then,

$$q'(t,\theta) > 0$$

and since

$$v(0,\theta) = v(1,\theta) = 0$$

q maps [0,1] monotonely onto itself. Let,

$$\tilde{g}_j(t,\theta) = g_j(q(t,\theta),\theta)q'(t,\theta). \tag{4.6}$$

If $g_j(\cdot,\theta)$ is the density of T given $Z = j$, then $\tilde{g}_j(\cdot,\theta)$ is the density of $q^{-1}(T,\theta)$. Note also that,

$$\tilde{g}_j(\cdot,\theta_0) = g_j.$$

The model corresponding to $\{\tilde{g}_j(\cdot,\theta), j = 0,1\}$ is a parametric submodel of the original model. Now

$$\frac{\partial}{\partial\theta}\log\tilde{g}_j(t,\theta) = \lambda_j(t,\theta) + \gamma_j(q(t,\theta),\theta)\sum_{l=0}^{1} v_l(t) + \frac{\partial}{\partial\theta}\log q'(t,\theta) \tag{4.7}$$

where,

$$\gamma_j(t,\theta) = -\frac{g_j'}{g_j}(t,\theta).$$

In particular,

$$\frac{\partial}{\partial\theta}\log\tilde{g}_j(t,\theta_0) = \lambda_j(t) - \Big(\sum_{l=0}^{1} v_l'(t) - \gamma_j(t)\sum_{l=0}^{1} v_l(t)\Big). \tag{4.8}$$

All functions in (4.7) are continuous in θ and uniformly bounded in t in a neighbourhood of θ_0. It follows that $\Lambda_N(\theta_N)$, the log-likelihood ratio of the data at $\theta_N = \theta_0 + tN^{-\frac{1}{2}}$ to θ_0, given $\underset{\sim}{Z}$ satisfies,

$$\Lambda_N(\theta_N) = \sum_{i=1}^{N}\sum_{j=0}^{1} Z_{ij}\log[\tilde{g}_j(T_i,\theta_N)/g_j(T_i)] \tag{4.9}$$

$$= N^{-\frac{1}{2}}\sum_{i=1}^{N}\sum_{j=0}^{1}\frac{\partial}{\partial\theta}\log\tilde{g}_j(T_i,\theta_0)Z_{ij} - \frac{t^2}{2}\sum_{j=0}^{1}\hat{\pi}_j\int_0^1\Big(\frac{\partial}{\partial\theta}\log\tilde{g}_j(u,\theta_0)\Big)^2 g_j(u)du$$

$$+ o_p(1)$$

under the pair (g_0,g_1). Under (A) it is also easy to check that,

$$\int_0^1 \lambda_j(t)g_j(t)dt = 0 \tag{4.10}$$

so that,

$$\sum_{i=1}^{N}\sum_{j=0}^{1} Z_{ij}\log\frac{\tilde{g}_j(T_i,\theta_N)}{g_j(T_i)} = N^{-\frac{1}{2}}t\sum_{i=1}^{N}\sum_{j=0}^{1} Z_{ij}A_j(T_i) \tag{4.11}$$

$$-\frac{t^2}{2}\sum_{j=0}^{1}\hat{\pi}_j\int_0^1 A_j^2(u)dG_j(u) + o_p(1).$$

Therefore, the test based on T_N is asymptotically most powerful for testing $H: \theta = \theta_0$ vs $K_N: \theta = \theta_N$ conditionally on $Z = j$, T has distribution $\tilde{g}_j(\cdot, \theta)$. Conditional efficiency of T_N in the sense of section 1 follows. If we take,

$$N\sigma_N^2 = \sum_{j=0}^{1} \hat{\pi}_j \int_0^1 A_j^2(u) dG_j(u)$$

then (1.3) holds with

$$a = \left(\sum_{j=0}^{1} \hat{\pi}_j \int_0^1 A_j^2(u) dG_j(u) \right)^{\frac{1}{2}}. \tag{4.12}$$

To see this note that by (4.11) and the boundedness of the A_j the measures induced by $\tilde{g}_j(\cdot, \theta_N)$ and g_j are contiguous. Again from (4.11) and (3.20),

$$\frac{tN^1}{2} T_N = \Lambda_N(\theta_N) + \frac{t^2}{2} N\sigma_N^2 + o_p(1). \tag{4.13}$$

Efficiency follows from this equivalence of T_N and the likelihood ratio test while (4.12) follows by a standard contiguity calculation e.g. Le Cam's third lemma, p. 208 Hajek and Sidak [10]. Since the conditional rank likelihood ratio test is at least as powerful as the test based on T_N its efficiency follows.

5. Extensions and Monte Carlo.

Extension of these results to the case Z finite say $= \{0, \ldots . p-1\}$ is straightforward. The only change needed in formula (1.14), (1.19), (2.1), (3.20), (3.21) etc. is to replace the upper limit of summation 1 by $p-1$. However, θ in such cases is typically multivariate so that one sided hypotheses without nuisance parameters are not very interesting. The extension to such hypotheses and estimation can be carried through by studying, under a fixed θ_0, the family of statistics $T_N(\theta)$, with λ_j, γ_j etc. chosen appropriate to θ being true, at least for $|\theta - \theta_0| = \mathcal{O}(N^{-1/2})$. Theorem 3.1 permits this kind of analysis. We intend to report on this subsequently.

It is relatively straightforward to show that for the Pareto family, including the Cox model for $c = 0$, γ_j is continuously twice differentiable on $[0,1)$ but,

$$\gamma_j^{(r)}(t) = \Omega((1-t)^{c-1-r}) \text{ as } t \to 1.$$

So,

$$\|\gamma_j'\|_1 = \infty \quad \text{if } c < 1. \tag{5.1}$$

For the normal model, α blows up and is not integrable at either 0 or 1. It appears that these difficulties can be resolved by considering statistics T_N based on a censored version of the data, $\underset{\sim}{Z}$ and $\{R_i : \epsilon_1 N \leq i \leq (1-\epsilon_2)N\}$ with $\epsilon_2, \epsilon_1 \downarrow 0$ at a slow enough rate. This analysis which is in progress should be extendable to the case of general right censoring and possibly also to time dependent covariates. Our results so far establish the efficiency of rank likelihood ratio tests. We expect that our extensions will show that estimation by maximizing the rank likelihood is generally efficient in transformation models, not just in the Cox model as was shown by Efron [9] and Begun et al [1]. This expected conclusion is supported by the results of Doksum [8]. Although we believe the approximation methods of Clayton and Cuzick which motivated our approach are similarly efficient we have not been able to analyze them successfully. How does our approach relate to the two situations for which formulae are known, (i) $f_0 = f_1$ and (ii) $f_0(t) = e^{-t}$, $f_1(t) = \theta e^{-\theta t}$, the Cox model? In (i) our approach leads to $g_0 = g_1 \equiv 1$ and hence $\gamma_j = \alpha \equiv 0$. So $\Delta(t, \hat{Q}_0, \hat{Q}_1) \equiv 0$ and we obtain one of the forms of the classical linear rank statistic for testing $H: f_0 = f_1$.

Our approach in (ii) requires censoring and in any case does not lead to an explicit form of T_N. However, if we follow the approach of Bickel [3] and do not reduce to (g_0, g_1) before linearizing (1.12) we arrive at an explicit quadratic rank statistic but based on a kernel which is of order

$(1-t)^{-1}$ near $t = 1$.

Monte Carlo The programming for this section was done by Julian Faraway to whom I am most grateful. We present some very limited Monte Carlo simulations.

We consider the Pareto model for $c = 1$ in the form

$$f_j(t,\theta) = e^{-(j-\frac{1}{2})\theta}(1+e^{-(j-\frac{1}{2})\theta}t)^{-2},\ j = 0,1$$

$$\textit{with} \quad m = n = 50,\ \theta_0 = 1.$$

We plot in figures 1-3 power curves, for levels of significance .1, .05, and .01 and statistics T_N, T_{N0}, T_{N1} where T_N is given by (1.19),

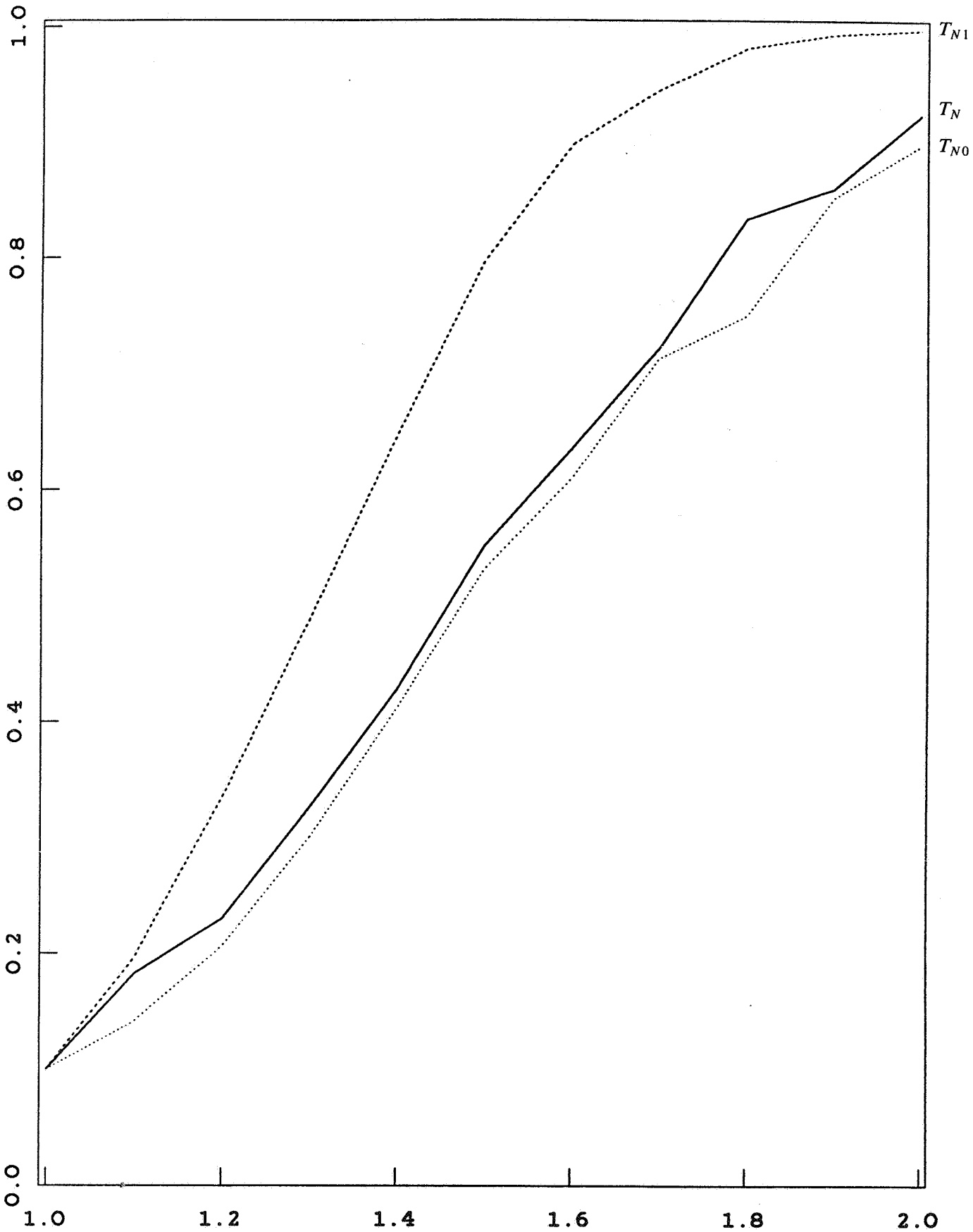
$$T_{N0} = \hat{\pi}_1 \int_0^1 \lambda(t) d\hat{Q}_1(t)$$

is the natural linear rank approximation to T_N while,

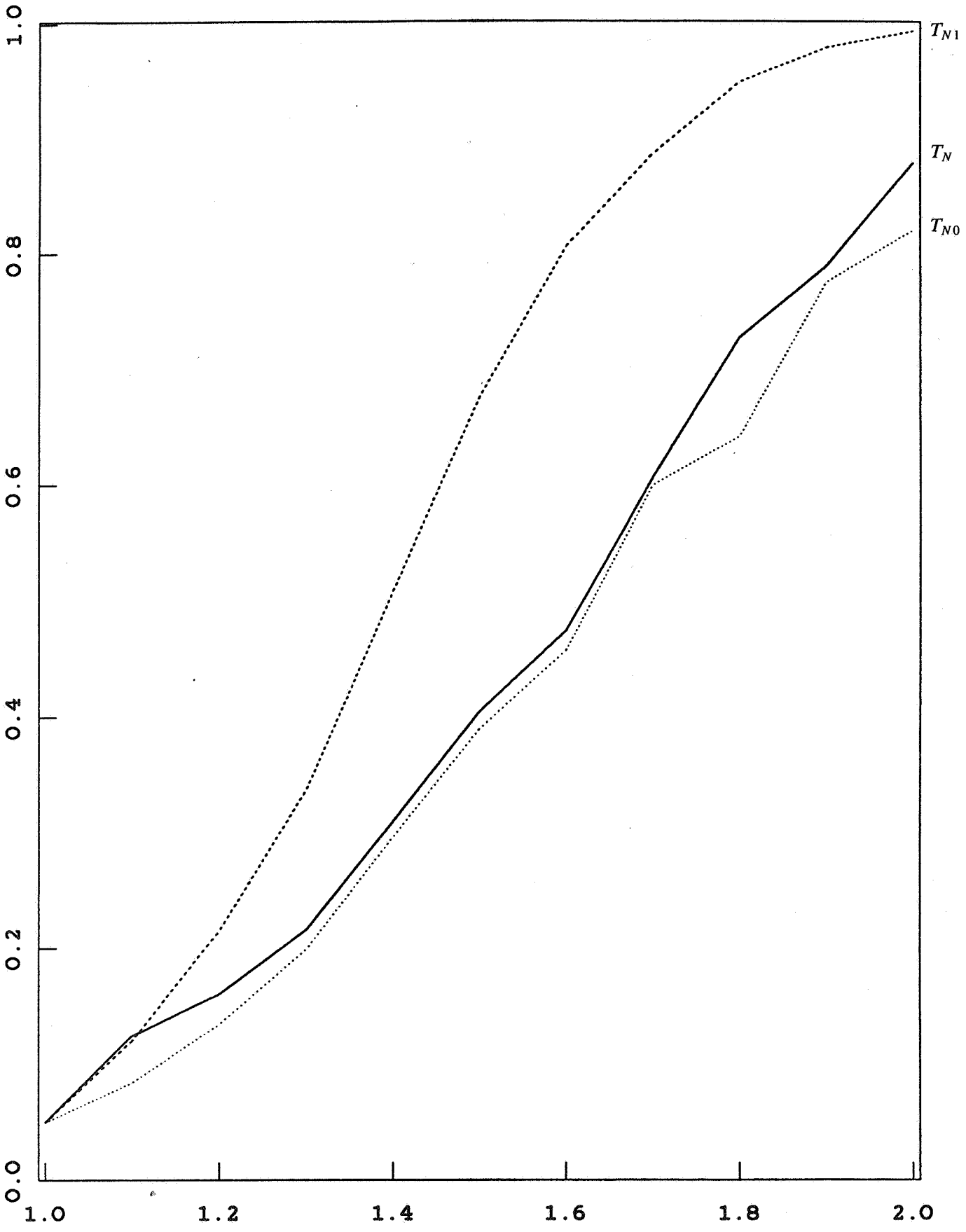
$$T_{N1} = \hat{\pi}_1 \int_0^1 e^{-\frac{\theta_0}{2}} t(1+e^{-\frac{\theta_0}{2}}t)^{-1} d\hat{G}_1(t) - \hat{\pi}_0 \int_0^1 e^{\frac{\theta_0}{2}} t(1+e^{\frac{\theta_0}{2}}t)^{-1} d\hat{G}_0(t)$$

is the locally most powerful test statistic for the parent family whose power is as we have seen unattainable by a rank statistic. Our results are based on 1000 simulations. The critical values were chosen using the simulations rather than the normal approximation as they could be in practice. Normal probability plots for T_N under the hypothesis show the normal approximation to be satisfactory.

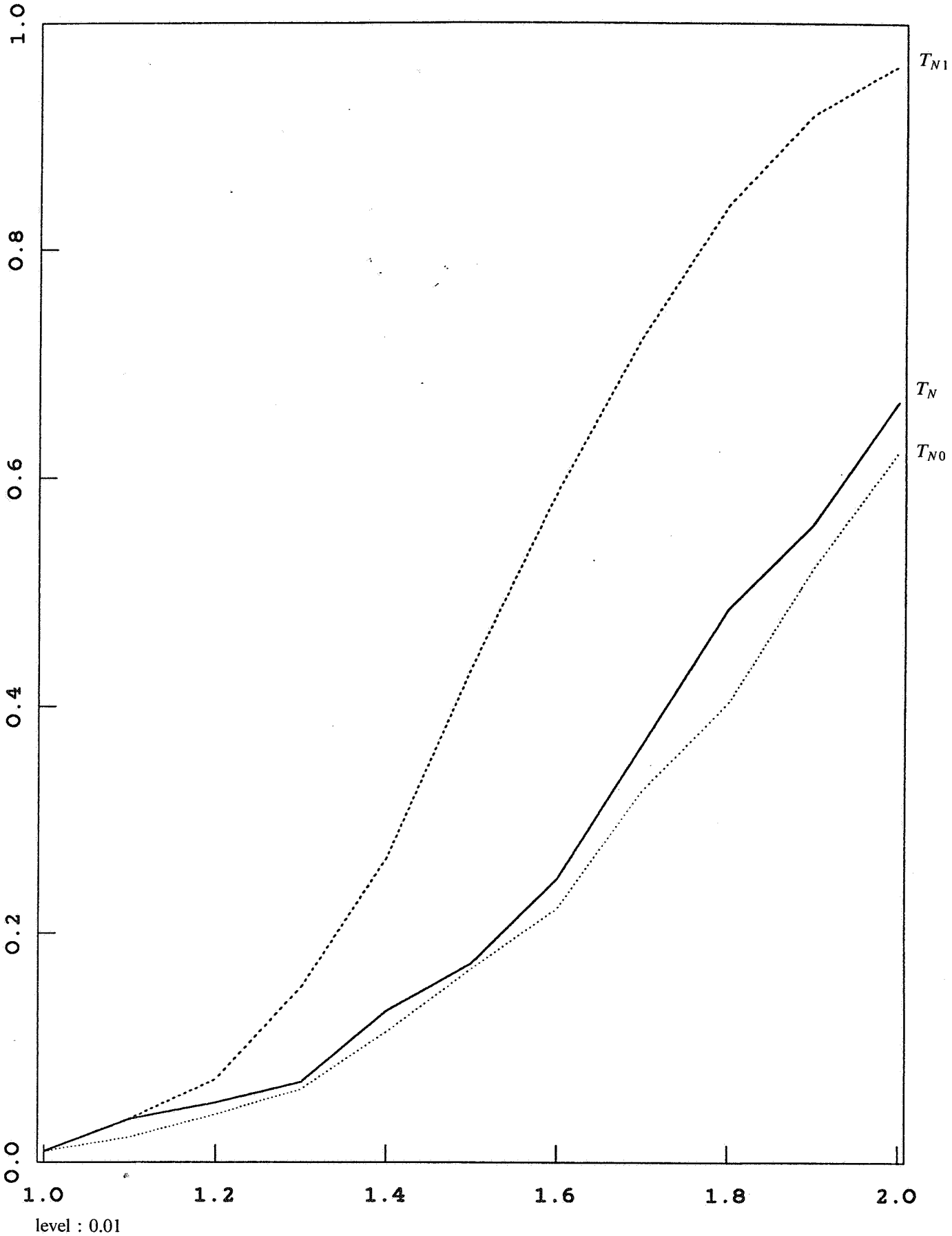
The power functions are consistently ordered as the theory predicts. T_N improves only slightly over T_{N0} for the situation considered. If this effect is observed more generally, we would advocate consistent use of the simpler linear rank approximation. Not knowing the transformation can cost a lot as measured by the discrepancy between T_{N1} and T_N. However, T_{N1} is the locally most powerful test for a model which is only an approximation (for θ_0 close to 0) to the actual least favorable model given by (4.6), so the comparison even in this case may be grossly unfair.



level : 0.10



level : 0.05



level : 0.01

REFERENCES

1. J.M. BEGUN, W.J. HALL, W.M. HUANG AND J.A. WELLNER (1983) Information and asymptotic efficiency in parametric and non-parametric models, *Ann. Statist. 11,* 432-452.
2. S. BENNETT (1983a). Log-logistic regression models for survival data. *Appl. Statist. 32,* 165-171.
3. P.J. BICKEL (1985). Discussion of papers on semiparametric models. *Proc.* I.S.I. Amsterdam; pp. 55-58 in this report.
4. D. CLAYTON AND J. CUZICK (1985). The semiparametric Pareto model for regression analysis of survival times. *Proc.* I.S.I. Amsterdam; pp. 19-30 in this report.
5. H. CHERNOFF AND I.R. SAVAGE (1958). Asymptotic normality and efficiency of certain non-parametric rank statistics *Ann. Math. Statist. 29,* 972-994.
6. D.R. COX (1972). Regression models and life-tables (with Discussion). *J. R. Statist. Soc. B, 34,* 187-220.
7. M. CSÖRGÖ (1983). Quantile processes with statistical applications CBMS-N.S.F. *Conference Series in Applied Mathematics 42,* S.I.A.M. Philadelphia.
8. K. DOKSUM (1985). (Preprint) *Partial likelihood methods in transformation models.*
9. B. EFRON (1977). The efficiency of Cox's likelihood function with censored data *J. Amer. Stat. Assoc. 72,* 557-565.
10. J. HAJEK AND Z. SIDAK (1967). *Theory of Rank Tests,* Academia, Prague.
11. E. HILLE (1969). *Lectures on ordinary differential equations.* Addison-Wesley, Reading Mass.
12. F. TRICOMI (1957). *Integral equations.* Academic Press N.Y.
13. R. PYKE AND G. SHORACK (1968). Weak convergence of a two sample empirical processes and a new approach to Chernoff-Savage theorems. *Ann. Math. Statist. 39,* 755-771.

50751
RP

MSC 62G05