A NOTE ON NONPARAMETRIC ESTIMATIONS

Péter Major[1] *and Lídia Rejtő,*[1,2]

[1] Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary

[2] Department of Mathematical Sciences, University of Delaware,
Newark, Delaware, USA

Dedicated to Miklós Csörgő on the occasion of his 65-th birthday

Abstract: We give an informal explanation with the help of a Taylor expansion about the most important properties of the maximum likelihood estimate in the parametric case. Then an analogous estimate in two nonparametric models, in the estimate of the empirical distribution function from censored data and in the Cox model is investigated. It is shown that an argument very similar to the proof in the parametric case yields analogous properties of the estimates in these cases too. There is an important non-trivial step in the proofs which is discussed in more detail. A double stochastic integral with respect to a standardized empirical process has to be estimated. This corresponds to the estimate of the second term of the Taylor expansion in the parametric case. We think that the method explained in this paper is applicable in several other models.

1. Some general remarks

An important problem of statistics is to estimate an unknown parameter or distribution by means of a sample of size n, i.e. by a sequence of independent and identically distributed random variables $\xi_1, \dots, \xi_n$ with some unknown distribution F. If this unknown distribution F belongs to a class of distribution functions $F(x, \vartheta)$, where ϑ is a real number or more generally the element of a finite dimensional vector space, and we are interested in the value ϑ or a function $g(\vartheta)$ of it, then we speak of parametric estimation. In the other case when the set of possible distributions where the sample can come from is an "infinite dimensional space" we speak of nonparametric estimation.

The case of parametric estimation is considerably simpler. In this case there is a powerful technique, the maximum likelihood method which has the following two nice properties:

i.) It supplies a method for a large class of problems.

ii.) Under general conditions it is asymptotically optimal.

In case of nonparametric estimation problems no such good and general method is available. Nevertheless, there are some special cases where such a good estimate can be given as in the parametric case. The investigation of both question, i.e.

i.) to find a general principle which enables us to give a good estimate in the nonparametric case

ii.) to show that the estimate is as good as the estimate in the parametric case

are challenging problems. There exists a large literature on nonparametric maximum likelihood estimation (see e.g. [bhhw], [gil1], [gil2], [lc]). Some special nonparametric models will be considered, and we prove that the estimates proposed in these cases are as good as the maximum likelihood estimate in parametric models. The structure of the proof is similar to that in the parametric case, but some additional technical problems appear. These problems and their solutions deserve special attention. To explain them first we present a short informal explanation about the limit behaviour of the maximum likelihood method in the parametric case. Here we assume that the distributions satisfy some natural smoothness conditions which we do not formulate explicitly.

Let us consider the simplest parametric problem when a parameter $\vartheta_0 \in \mathbf{R}^1$ has to be estimated from a class of distribution functions $F(x,\vartheta)$, $\vartheta \in \mathbf{R}^1$, by means of a sequence of independent random variables $\xi_1(\omega),\dots,\xi_n(\omega)$ with distribution $F(x,\vartheta_0)$.

We also assume that the distribution functions $F(x,\vartheta)$ have a density function $f(x,\vartheta)$ with respect to a measure μ on the real line. The maximum likelihood method suggests to choose the estimate $\hat\vartheta_n = \hat\vartheta_n(\xi_1,\dots,\xi_n)$ of the parameter ϑ_0 as the number where the density function of the random vector $(\xi_1,\dots,\xi_n)$ (with respect to the product measure $\underbrace{\mu\times\cdots\times\mu}_{n \text{ times}}$), i.e. the product

$$\prod_{k=1}^n f(\xi_k,\vartheta) = \exp\left\{\sum_{k=1}^n \log f(\xi_k,\vartheta)\right\}$$

takes its maximum. This point can be found as the solution of the equation

$$\sum_{k=1}^n \frac{\partial}{\partial\vartheta}\log f(\xi_k,\vartheta) = 0\,. \tag{1.1}$$

We are interested in the asymptotic behaviour of the random variable $\hat\vartheta_n - \vartheta_0$, where $\hat\vartheta_n$ is the (appropriate) solution of the equation (1.1). Let us take Taylor expansion of the expression at the left hand side of (1.1) around the point ϑ_0. We get

$$\begin{aligned}\sum_{k=1}^n \frac{\partial}{\partial\vartheta}\log f(\xi_k,\hat\vartheta_n) &= \sum_{k=1}^n \frac{\frac{\partial}{\partial\vartheta}f(\xi_k,\vartheta_0)}{f(\xi_k,\vartheta_0)} \\ &\quad + (\hat\vartheta_n-\vartheta_0)\left(\sum_{k=1}^n\left(\frac{\frac{\partial^2}{\partial\vartheta^2}f(\xi_k,\vartheta_0)}{f(\xi_k,\vartheta_0)} - \frac{\left(\frac{\partial}{\partial\vartheta}f(\xi_k,\vartheta_0)\right)^2}{f^2(\xi_k,\bar\vartheta_0)}\right)\right) \\ &\quad + O\left(n(\hat\vartheta_n-\vartheta_0)^2\right) \\ &= \sum_{k=1}^n\Big(\eta_k + \zeta_k(\hat\vartheta_n-\vartheta_0)\Big) + O\left(n(\hat\vartheta_n-\vartheta_0)^2\right)\,,\end{aligned} \tag{1.2}$$

where

$$\eta_k = \frac{\frac{\partial}{\partial\vartheta} f(\xi_k, \vartheta_0)}{f(\xi_k, \vartheta_0)} \quad \text{and} \quad \zeta_k = \frac{\frac{\partial^2}{\partial\vartheta^2} f(\xi_k, \vartheta_0)}{f(\xi_k, \vartheta_0)} - \frac{\left(\frac{\partial}{\partial\vartheta} f(\xi_k, \vartheta_0)\right)^2}{f^2(\xi_k, \bar{\vartheta}_0)}$$

for $k = 1, \dots, n$. We want to understand the asymptotic behaviour of the (random) expression on the right-hand side of (1.2). The relation

$$E\eta_k = \int \frac{\frac{\partial}{\partial\vartheta} f(x, \vartheta_0)}{f(x, \vartheta_0)} f(x, \vartheta_0)\, d\mu(x) = \frac{\partial}{\partial\vartheta} \int f(x, \vartheta_0)\, d\mu(x) = 0$$

holds, since $\int f(x, \vartheta)\, d\mu(x) = 1$ for all ϑ, and differentiating this relation we get the last identity. Similarly, $E\eta_k^2 = -E\zeta_k = \int \frac{\left(\frac{\partial}{\partial\vartheta} f(x,\vartheta_0)\right)^2}{f(x,\vartheta_0)}\, d\mu(x) > 0, \quad k = 1, \dots, n$. Hence by the central limit theorem

$$\chi_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \eta_k$$

is asymptotically normal with expectation zero and variance

$$I^2 = \int \frac{\left(\frac{\partial}{\partial\vartheta} f(x, \vartheta_0)\right)^2}{f(x, \vartheta_0)}\, d\mu(x) > 0\,.$$

In the statistics literature this number I is called the Fisher information. By the laws of large numbers $\frac{1}{n} \sum_{k=1}^{n} \zeta_k \sim -I^2$. Hence it follows from relation (1.2) that

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{\partial}{\partial\vartheta} \log f(\xi_k, \hat{\vartheta}_n) = \chi_n - \sqrt{n}(\hat{\vartheta}_n - \vartheta_0)\, I^2 + \text{negligible error.} \tag{1.3}$$

Formulas (1.1) and (1.3) imply that

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) = I^{-2}\, \chi_n + \text{negligible error,} \tag{1.4}$$

which is asymptotically a normal random variable with expectation zero and variance I^{-2}. Another result of the mathematical statistics, the so-called Cramer–Rao inequality states that this result is asymptotically optimal, since under general conditions all unbiased estimates (such estimates whose expectation equals the estimated parameter) have a variance which multiplied by $\sqrt{n}$ cannot be smaller than I^{-2}.

We are also interested in the magnitude of the "negligible error" term in formula (1.4) and want to compare it with the error term appearing in the analogous formulas proved for the nonparametric estimates discussed in Sections 2 and 3. We shall show that the (normalized) error can be approximated by (normalized) sums of independent random variables in the parametric and by a linear functional of the standardized empirical distribution function in the nonparametric case. We give an informal explanation that the error of this approximation is of order $O(n^{-1/2})$ in both cases.

In this section we discuss the parametric case. Put

$$\varepsilon_n = \sqrt{n}\left(\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) - I^{-2}\chi_n\right)$$

and

$$\omega_n = n^{-1/2}\sum_{k=1}^{n}(\zeta_k + I^2)\ ,$$

and express $\sqrt{n}(\hat{\vartheta}_n - \vartheta_0)$ and $\sum_{k=1}^{n}\zeta_k$ through ε_n and ω_n. Then relations (1.1) and (1.2) imply that

$$\sqrt{n}\chi_n + (n^{-1/2}\varepsilon_n + I^{-2}\chi_n)(\omega_n - \sqrt{n}I^2) + O\left((n^{-1/2}\varepsilon_n + I^{-2}\chi_n)^2\right) = 0\ ,$$

or equivalently

$$\omega_n\chi_n I^{-2} - I^2\varepsilon_n + n^{-1/2}\omega_n\varepsilon_n + O\left((n^{-1/2}\varepsilon_n + I^{-2}\chi_n)^2\right) = 0\ .$$

Since the random variables χ_n and ω_n are stochastically bounded, the last relation implies that $\varepsilon_n = \sqrt{n}\left(\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) - I^{-2}\chi_n\right)$ is stochastically bounded. This is equivalent to saying that the normalized error $\sqrt{n}(\hat{\vartheta}_n - \vartheta_0)$ of the maximum likelihood estimate can be approximated by an appropriate normalized sum of independent random variables (by $I^{-2}\chi_n$) with a (random) error of order $n^{-1/2}$. A longer Taylor expansion in formula (1.2) and an Edgeworth expansion for sums of independent random variables yields a better approximation, an Edgeworth type expansion for $\sqrt{n}(\vartheta - \vartheta_0)$.

We are interested in what can be preserved from the methods and results of the parametric estimates when nonparametric estimation problems are considered. Since in nonparametric models typically the set of all probability measures in the model cannot have a density function with respect to a fixed measure, the maximum likelihood principle cannot be applied. Nevertheless, in several interesting models a good estimate can be found by an appropriate modification of the maximum likelihood argument. We shall consider such models and show that these estimates are good.

The simplest nonparametric estimation problem is the estimation of an unknown distribution function $F(x)$ by means of a sequence of independent F distributed random variables $\xi_1, \dots, \xi_n$. Define the empirical distribution function

$$F_n(x) = \frac{1}{n}\#\{k:\ \xi_k \le x\}\ .$$

By a classical result of probability theory the processes $\sqrt{n}(F_n(x) - F(x))$ converge weakly to a Gaussian process $Z(x)$, $-\infty < x < \infty$, with expectation zero an covariance function $EZ(x)Z(y) = \min\{F(x), F(y)\} - F(x)F(y)$. We want to show that a similar limit theorem holds for $\sqrt{n} \times$ (the estimate $-$ the real value) for a good estimate in other estimation problems. Moreover, we want to investigate the distance of this expression for fixed n from the limit process. In the parametric case this expression

can be approximated by the normalized sum of independent random variables, or what is equivalent to this, by a normal random variable with an accuracy of order $n^{-1/2}$. Actually, it would demand a more detailed explanation of how we measure the accuracy of approximation in the last statement. But we shall not do this, we only explain the content of this statement in the nonparametric models discussed in the sequel.

We shall discuss two models, the so-called random censorship and the Cox model and write down an estimate in both models and show that they are good. The difference of the estimate and the real distribution multiplied by $\sqrt{n}$ converges to a Gaussian process. Moreover, this difference can be approximated by a linear functional of a normalized empirical process with an error of order $n^{-1/2}$. More explicitly, the error is a random process, and if it is multiplied by $\sqrt{n}$ then the probability that the supremum of this process is larger than some $x > 0$ is smaller than $C_1 e^{-\lambda x}$ with some constants C and λ independent of n and x.

The linear functional of a standardized empirical function is the natural infinite dimensional counterpart of sums of independent random variables which appears in the parametric case. In the nonparametric estimates considered in this paper the main contribution to the error (multiplied by $\sqrt{n}$) of the estimate is expressed in such a form. It converges to a Gaussian process as the sample size $n \to \infty$. The difference of the estimate and the real distribution multiplied by $\sqrt{n}$ can also be approximated by an appropriate Gaussian process. This appears as the limit of the linear functionals of the empirical processes that yield the main contribution to the (normalized) error of the estimate. The goodness of this Gaussian approximation can be determined with the help of the bound on the linear functional approximation of the error and a result of Komlós, Major and Tusnády [kmt] (see in [csr] for more details) about the approximation of the empirical process by a Brownian bridge.

The proof of these results is based on an expansion which can be considered as an adaptation of the investigation of the maximum likelihood estimate in the parametric case. One point of the proof deserves special attention. If we want to adapt the method of the parametric case, then an expression analogous to formula (1.2) has to be studied. A term which corresponds to the error term $O\left(n(\hat{\vartheta}_n - \vartheta)^2\right)$ of the Taylor expansion in (1.2) has to be well estimated. In the nonparametric case this problem can be solved if the distribution of certain non-linear functionals of the empirical distribution function of the sample $(\xi_1, \dots, \xi_n)$ is well bounded.

Hence the following problem has to be studied. Let us consider the normalized empirical distribution of a sample, the s-fold direct product of their (random) measure with itself, and let us estimate the integral of a bounded function of s variables with respect to this product measure. If $s \geq 2$, then this integral is a non-linear functional of the empirical distribution. We are mainly interested in the special case $s = 2$, but the results will be formulated for general s. The proof of these results can be found in [mp], [mr], and [mrsi]. First we formulate the result in a slightly restrictive form, when the random variables whose empirical distribution is considered are uniformly distributed in the interval $[0, 1]$. This result is sufficient for instance in the investigation of the random censorship model.

Theorem A. *Let $\xi_1(\omega), \dots, \xi_n(\omega)$ be independent uniformly distributed random vari-*

ables on $[0,1]$, $F_n(u) = F_n(u,\omega) = \frac{1}{n}\#\{k : 1 \le k \le n,\ \xi_k \le u\}$, $0 \le u \le 1$, *their empirical distribution function and* $\mu_n(u) = \sqrt{n}(F_n(u) - u)$ *the standardization of this empirical distribution function. Let* $f(u_1, \dots, u_s)$ *be a function on* $[0,1]^s$ *such that* $\sup_{u_1,\dots,u_s} |f(u_1, \dots, u_s)| \le 1$, *and* $f(u_1, \dots, u_s) = 0$ *if* $u_j = u_k$ *with some* $1 \le j < k \le s$. *There exist some universal constants* C_s *and* α_s *depending only on the dimension* s *in such a way that*

$$P\left(\sup_{0\le t\le 1}\left|\int_0^t\int_0^1\cdots\int_0^1 f(u_1,\dots,u_s)d\mu_n(u_1)\dots d\mu_n(u_s)\right| \ge x\right) \le C_s \exp\left\{-\alpha_s x^{2/s}\right\}$$

for all $x > 0$, *and function* f *with the above properties.*

In other cases, like in the Cox model, the integral with respect to the product measure of an empirical distribution in a higher dimensional Euclidean space is needed. The proof of such results (actually the reduction of such results to the case described in Theorem A) is not harder in the case of general separable metric spaces, hence we formulate the result in such a form. To do this we introduce some notations.

Let a probability space $(\Omega, \mathcal{A}, P)$ and a separable metric space $(X, \mathcal{X})$ be given. Let $\xi\colon \Omega \to X$ be an X valued random variable on $(\Omega, \mathcal{A}, P)$. Let μ denote the distribution of the random variable ξ, i.e. let

$$\mu(B) = P(\xi \in B) = P(\xi^{-1}(B)) \qquad \forall\, B \in \mathcal{X}\ .$$

Suppose that $\xi_1, \xi_2, \dots, \xi_n$ are independent, identically distributed random variables on $(\Omega, \mathcal{A}, P)$ with values on the space $(X, \mathcal{X})$ and distribution μ. We introduce the empirical measure

$$\bar{\mu}_n(B) = \frac{1}{n}\sum_{i=1}^n I(\xi_i \in B) \qquad \forall\, B \in \mathcal{X}\ ,$$

and its standardization

$$\mu_n(B) = \sqrt{n}\,(\bar{\mu}_n(B) - \mu(B)) \qquad \forall\, B \in \mathcal{X}\ .$$

Let X_t, $0 \le t \le 1$ be a system of sets in $\mathcal{X}$ with the following property:

Property (i) $X_s \subseteq X_t$ for all $s \le t$, $X_0 = \emptyset$, $X_1 = X$, $\mu(X_t) = t$.

Let us consider the product space $\underbrace{X \times \cdots \times X}_{s \text{ times}} = X^s$ with product measure $\mu^{(s)}(\cdot)$ and the diagonal set $A \in X^s$ is defined as

$$A = \{(x_1, \dots x_s)\colon x_i = x_j \quad \text{for some } i \ne j\}$$

Let $\mathcal{F}$ denote the set of the real valued measurable functions $f(u_1, \dots, u_s)$ defined on the space X^s whose absolute value is less than 1, and which disappear on the diagonal set A, i.e. let

$$\mathcal{F} = \{f(u_1, \dots, u_s)\colon |f| \le 1, \quad f(u_1, \dots, u_s) = 0\ \forall\ (u_1, \dots, u_s) \in A\}\ .$$

Then

Theorem B. *There exist some universal constants C_s and α_s depending only on the dimension s in such a way that*

$$P\left(\sup_{0\le t\le 1}\left|\int_{X_t}\int_X\cdots\int_X f(u_1,\dots,u_s)d\mu_n(u_1)\dots d\mu_n(u_s)\right|\ge x\right)\le C_s\exp\left\{-\alpha_s x^{2/s}\right\}$$

for all $f\in\mathcal{F}$ and $x>0$, where the sets X_t, $0\le t\le 1$ satisfy Property (i).

Theorem B shows certain analogy with respect to multiple stochastic integrals with respect to a Gaussian process. Here the underlying process, the empirical distribution function is not Gaussian, but it is almost Gaussian. In both cases the diagonal is cut from the domain of integration. The (random) measures of disjoint intervals are almost independent. In Theorem B actually we investigate how strong cancellation is caused by this almost independence. The upper bound $C\exp\left\{-\lambda x^{2/s}\right\}$ given for the tail distribution of an s-fold integral is sharp. It expresses the fact that the tail behaviour of an s-fold stochastic integral is similar to the tail behaviour of the distribution of the s-th power of a Gaussian random variable.

In the statistical problems we discuss below, we first write up the statistics under investigation as a multiple integral with respect to an empirical measure plus some possible additional terms we can handle. Then we have to handle an expression of the form

$$Z_n(t)=\int_{X_t}\int_X g(u_1,u_2)\,d\bar\mu_n(u_1)\,d\bar\mu_n(u_2)\,, \tag{1.5}$$

where $\bar\mu_n$ denotes the empirical distribution (without normalization). With the help of simple algebra we get that

$$\begin{aligned}d\bar\mu_n(u_1)\,d\bar\mu_n(u_2)&=d\mu(u_1)\,d\mu(u_2)+\frac{1}{\sqrt{n}}d\mu(u_1)d\mu_n(u_2)\\&\quad+\frac{1}{\sqrt{n}}d\mu(u_1)d\mu_n(u_2)+\frac{1}{n}d\mu_n(u_1)\,d\mu_n(u_2)\,,\end{aligned}$$

where $\mu_n(\cdot)=\sqrt{n}(\bar\mu_n(\cdot)-\mu(\cdot))$. Then we can write

$$Z_n(t)=D(t)+n^{-1/2}U_n(t)+n^{-1}V_n(t)\,, \tag{1.6}$$

where

$$D(t)=\int_{X_t}\int_X g(u_1,u_2)\,d\mu(u_1)\,d\mu(u_2)$$

is a deterministic function,

$$U_n(t)=\int_{X_t}H_1(u)\,d\mu_n(u)+\int_X H_2(u)\,d\mu_n(u)$$

with $H_1(u)=\int_X g(u,u_2)\,d\mu(u_2)$ and $H_2(u)=\int_{X_t}g(u_1,u)d\mu(u_1)$ is a linear functional of the empirical distribution μ_n and the term $V_n(t)$ is a non-linear function of the

measure μ_n. Because of Theorem B we have a good bound on the distribution of $\sup_{0\le t\le 1} |V_n(t)|$. Hence we can handle the expression

$$\sqrt{n}(Z_n(t) - D(t)) = U_n(t) + \frac{1}{\sqrt{n}} V_n(t) \ .$$

In the following examples we shall see why this observation is useful in the study of certain nonparametric estimates.

2. The Kaplan-Meier product limit estimator

In this section the following problem is considered. Let (X_i, C_i), $i = 1, \dots, n$, be a sequence of independent, identically distributed random vectors such that the components X_i and C_i are also independent with distribution functions $F(x)$ and $G(x)$. We want to estimate the distribution function F of the random variables X_i, but we cannot observe the variables X_i, only the random variables $Y_i = \min(X_i, C_i)$ and $\delta_i = I(X_i \le C_i)$. For the sake of simplicity we assume that both distributions F and G have no atom. In other words, we want to solve the following problem. There are certain objects whose lifetime X_i are independent and F distributed. But we cannot observe this lifetime X_i, because after a time C_i the observation must be stopped. We also know whether the real lifetime X_i or the censoring variable C_i was observed. We make n independent experiments and want to estimate with their help the distribution function F.

It is not easy to find the right estimate of the distribution function F on the basis of the above observations. Kaplan and Meier in [km] proposed the so-called product limit estimator to estimate the unknown survival function $S = 1 - F$ on the basis of the above observations. They proposed the following estimator S_n:

$$1 - F_n(u) = S_n(u) = \begin{cases} \displaystyle\prod_{i=1}^{n} \left(\frac{N(Y_i)}{N(Y_i)+1} \right)^{I(Y_i \le u, \delta_i = 1)} & \text{if } u \le \max(Y_1, \dots, Y_n) \\ 0 & \text{if } u \ge \max(Y_1, \dots, Y_n),\ \delta_n = 1, \\ \text{undefined} & \text{if } u \ge \max(Y_1, \dots, Y_n),\ \delta_n = 0, \end{cases} \tag{2.1}$$

where

$$N(t) = \#\{Y_i,\ \ Y_i > t,\ 1 \le i \le n\} = \sum_{i=1}^{n} I(Y_i > t) \ .$$

One would like to understand how to find the estimate (2.1) and why it is good. We do not discuss the first question, we only refer to some papers where the answer to this question is explained. The estimator (2.1) is a generalized maximum likelihood estimator (GMLE) of the unknown distribution function. The meaning of GMLE needs an explanation because in the nonparametric case we are searching for a measure in a non-dominated family of probability measures. Kiefer and Wolfowitz [kw] gave a definition for GMLE (see [rt] in this volume), and Johansen [joh] proved that the product limit estimator is GMLE in this sense.

We want to show that the estimate (2.1) is really good. This expression in its original form is rather complicated and hard to study. But it can be rewritten in a form more appropriate for our purposes. We briefly explain how this can be done. Our calculation leading to a better representation of the expression (2.1) closely follows the paper [mr] where the details are worked out. We give an expansion of the random variable S_n defined in (2.1). The leading term of this expansion is $1-F(u)$, the quantity we wanted to estimate. The second term is $n^{-1/2}$ times a linear functional of a standardized empirical distribution function, and the remaining error term can be bounded by n^{-1} times a random variable with finite moment generating function. In such a way a result can be proved which shows that the Kaplan–Meier estimate behaves very similarly to the maximum likelihood estimate in the parametric case. The method of the proof deserves special attention since it is also applicable in case of other nonparametric GMLE-s.

In the calculations, similarly to the study of the maximum likelihood estimate, appropriate Taylor expansions can be made, and the error term of these expansions can be well bounded. Most steps are routine, but there is a step which deserves special attention. During our calculations we have to estimate a quadratic functional of a standardized empirical distribution function, and this estimate is non-trivial. This corresponds to the estimate of the second term of the Taylor expansion in the maximum likelihood estimate in the parametric case, and such an expression can be well bounded by the result formulated in Theorem B of this paper.

First we introduce some notations. Put

$$\begin{aligned} &H(u) = P(Y_i \le u) = 1 - \bar{H}(u), \\ &\tilde{H}(u) = P(Y_i \le u,\, \delta_i = 1), \quad \tilde{\tilde{H}}(u) = P(Y_i \le u,\, \delta_i = 0), \end{aligned} \tag{2.2}$$

and

$$\begin{aligned} &H_n(u) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le u), \\ &\tilde{H}_n(u) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le u,\, \delta_i = 1), \quad \tilde{\tilde{H}}_n(u) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le u,\, \delta_i = 0)\ . \end{aligned} \tag{2.3}$$

Clearly $H(u) = \tilde{H}(u) + \tilde{\tilde{H}}(u)$ and $H_n(u) = \tilde{H}_n(u) + \tilde{\tilde{H}}_n(u)$. We consider $F_n(u) - F(u)$ on an interval $(-\infty, T]$, where

$$1 - H(T) > \delta \quad \text{with some fixed } \delta > 0\ . \tag{2.4}$$

We introduce the so-called cumulative hazard function and its empirical version

$$\Lambda(u) = -\log(1 - F(u)), \quad \Lambda_n(u) = -\log(1 - F_n(u))\ . \tag{2.5}$$

From (2.1) it is obvious that

$$\Lambda_n(u) = -\sum_{i=1}^{n} I(Y_i \le u,\, \delta_i = 1) \log\left(1 - \frac{1}{1 + N(Y_i)}\right).$$

Since $F_n(u) - F(u) = \exp(-\Lambda(u))\,(1 - \exp(\Lambda(u) - \Lambda_n(u)))$ a simple Taylor expansion yields

$$F_n(u) - F(u) = (1 - F(u))\,(\Lambda_n(u) - \Lambda(u)) + R_1(u)\,, \tag{2.6}$$

and it is easy to see that $R_1(u) = O\left(\Lambda(u) - \Lambda_n(u)\right)^2\big)$.

It follows from the subsequent estimations that $\sup\limits_{u \le T} \sqrt{n}|\Lambda(u) - \Lambda_n(u)|$ has exponential tail, thus the same is true for $\sup\limits_{u \le T} n|R_1(u)|$. Hence it is enough to investigate the term $\Lambda_n(u) - \Lambda(u)$.

We approximate $\Lambda_n(u)$ with the help of the relation $-\log(1-x) = x + O(x^2)$ for small x. We get

$$\Lambda_n(u) = \sum_{i=1}^{n} \frac{I(Y_i \le u,\, \delta_i = 1)}{N(Y_i)} + R_2(u) = \tilde{\Lambda}_n(u) + R_2(u)\,, \tag{2.7}$$

and the error term $nR_2(u)$ has also exponential tail (e.g. [mr] for the details).

The expression $\tilde{\Lambda}_n(u)$ is still not appropriate for our purposes. Since the denominators $N(Y_i) = \sum\limits_{j=1}^{n} I(Y_j > Y_i)$ are dependent on different i's, we cannot see directly the limiting behaviour of $\tilde{\Lambda}_n(u)$.

By exploiting the fact that the conditional distribution of $N(Y_i)$ given Y_i is a binomial distribution with parameters $n-1$ and $1 - H(Y_i)$, we can rewrite $\tilde{\Lambda}_n(u)$ in a more appropriate form. We shall approximate it by an expression which can be handled better. By writing

$$N(Y_i) = \sum_{j=1}^{n} I(Y_j > Y_i) = n\bar{H}(Y_i) \left(1 + \frac{\sum\limits_{j=1}^{n} I(Y_j > Y_i) - n\bar{H}(Y_i)}{n\bar{H}(Y_i)} \right),$$

and applying the inequality $\left| \dfrac{1}{1+z} - 1 + z \right| < 2z^2$, for $|z| < \dfrac{1}{2}$, with the choice $z = \dfrac{\sum\limits_{j=1}^{n} I(Y_j > Y_i) - n\bar{H}(Y_i)}{n\bar{H}(Y_i)}$, we obtain

$$\begin{aligned} \tilde{\Lambda}_n(u) &= \sum_{i=1}^{n} \frac{I(Y_i \le u,\, \delta_i = 1)}{n\bar{H}(Y_i)} \left(1 - \frac{\sum\limits_{j=1}^{n} I(Y_j > Y_i) - n\bar{H}(Y_i)}{n\bar{H}(Y_i)} \right) + R_3(u) \\ &= 2A(u) - B(u) + R_3(u), \end{aligned} \tag{2.8}$$

where

$$A(u) = A(n,u) = \sum_{i=1}^{n} \frac{I(Y_i \le u,\, \delta_i = 1)}{n\bar{H}(Y_i)}$$

and

$$B(u) = B(n,u) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{I(Y_i \le u,\, \delta_i = 1) I(Y_j > Y_i)}{n^2 \bar{H}^2(Y_i)} \ .$$

Again the reader is referred to [mr] for the tail behaviour of $nR_3(u)$. Thus (2.7) and (2.8) together yield

$$\Lambda_n(u) = 2A(u) - B(u) + \text{negligible error}, \tag{2.9}$$

and the sums A and B can be rewritten as stochastic integrals in the same way as in [mr]. Finally one obtains

$$\begin{aligned} \sqrt{n}\,(\Lambda_n(u) - \Lambda(u)) = {} & \frac{\sqrt{n}\left(\tilde{H}_n(u) - \tilde{H}(u)\right)}{1 - H(u)} - \int_{-\infty}^{u} \frac{\sqrt{n}(\tilde{H}_n(y) - \tilde{H}(y))}{(1 - H(y))^2}\, dH(y) \\ & + \int_{-\infty}^{u} \frac{\sqrt{n}\,(H_n(y) - H(y))}{(1 - H(y))^2}\, d\tilde{H}(y) \\ & - \sqrt{n} B_1(u) + \text{negligible error}, \end{aligned} \tag{2.10}$$

where

$$B_1(u) = \frac{1}{n} \int_{-\infty}^{u} \int_{-\infty}^{+\infty} \frac{I(x > y)}{(1 - H(y))^2}\, d\left(\sqrt{n}\,(H_n(x) - H(x))\right)\, d\left(\sqrt{n}(\tilde{H}_n(y) - \tilde{H}(y))\right) \ .$$

This formula is the analogous one of (1.6). To prove this we still have to show that the term $B_1(u)$ is also small. Theorem A suggests such an estimate. However, this result cannot be applied directly in the present case, since in the integral defining B_1, one has to integrate with respect to two different processes in the variables x and y. In the paper [mr] we could overcome this difficulty by rewriting B_1 as a double integral of an appropriate kernel function with respect to a standardized empirical process which contains all information on $H_n(\cdot)$ and $\tilde{H}_n(\cdot)$. Here we choose a different argument. We deduce the needed estimate directly from Theorem B. The advantage of this argument is that it is more flexible and applicable in other cases too. Instead of using H, H_n and $\tilde{H}$, $\tilde{H}_n$, we use the two dimensional measure $\mu(x_1, x_2) = F(x_1)G(x_2)$ and the empirical measure

$$\bar{\mu}_n(x_1, x_2) = \frac{1}{n} \#\{i \colon 1 \le i \le n,\ X_i \le x_1, C_i \le x_2\}$$

on $\mathbf{R}^2$. Since they contain all information, the expression $B_1(u)$ can be rewritten as a double stochastic integral with respect to the measure $\mu_n = \sqrt{n}(\bar{\mu}_n - \mu)$. To see this observe that

$$\tilde{H}([a,b]) = \mu\left(\mathbf{A}([a,b])\right), \quad \tilde{H}_n([a,b]) = \bar{\mu}_n\left(\mathbf{A}([a,b])\right),$$

$$\tilde{\tilde{H}}([a,b]) = \mu\left(\mathbf{B}([a,b])\right), \quad \tilde{\tilde{H}}_n([a,b]) = \bar{\mu}_n\left(\mathbf{B}([a,b])\right) \ ,$$

where

$$\mathbf{A}([a,b]) = \{(x_1, x_2) \colon a \le x_1 \le b,\ x_1 \le x_2\},$$

and

$$\mathbf{B}([a,b]) = \{(x_1,x_2)\colon a \le x_2 \le b,\ x_2 \le x_1\}\ .$$

Then applying the decompositions $H = \tilde{H} + \tilde{\tilde{H}}$ and $H_n = \tilde{H}_n + \tilde{\tilde{H}}_n$, we obtains

$$B_1(u) = \frac{1}{n} \int\limits_{y_1<u} \int\limits_{\mathbf{R}^2} \frac{I(x_1 > y_1, x_1 \le x_2, y_1 \le y_2)}{(1-H(y_1))^2}\, d\mu_n(x_1,x_2)\, d\mu_n(y_1,y_2)$$
$$+ \frac{1}{n} \int\limits_{y_1<u} \int\limits_{\mathbf{R}^2} \frac{I(x_1 > y_1, x_2 \le x_1, y_1 \le y_2)}{(1-H(y_1))^2}\, d\mu_n(x_1,x_2)\, d\mu_n(y_1,y_2)\ .$$

Then Theorem B makes possible to estimate $\sup\limits_{u\le T} |B_1(u)|$ if the number T satisfies (2.4), since the integrand is bounded in this case. We omit the details and we only formulate the final result we get in such a way. A detailed proof can be found in [mr].

Theorem 2.1 *Let T be such that $1 - H(T) > \delta$ with some $\delta > 0$. Then the process $F_n(u) - F(u)$, $-\infty < u < T$, where $1 - F_n(u)$ is defined in formula (2.1) can be represented as*

$$F_n(u) - F(u) = (1-F(u))(U(n,u) + V(n,u)) + R(n,u)\ , \quad -\infty < u < T\ ,$$

where

$$\sqrt{n}U(n,u) = \frac{\sqrt{n}(\tilde{H}_n(u) - \tilde{H}(u))}{(1-H(u))} - \int\limits_{-\infty}^{u} \frac{\sqrt{n}(\tilde{H}_n(y) - \tilde{H}(y))}{(1-H(y))^2}\, dH(y)$$

$$\sqrt{n}V(n,u) = \int\limits_{-\infty}^{u} \frac{\sqrt{n}(H_n(y) - H(y))}{(1-H(y))^2}\, dH(y)$$

are linear functionals of the empirical processes $\sqrt{n}(H_n(y) - H(y))$ and $\sqrt{n}(\tilde{H}_n(u) - \tilde{H}(u))$, and the error term $R(n,u)$ can be bounded as

$$P\left(\sup_{u\le T} n|R(n,u)| > x + \frac{C}{\delta}\right) \le Ke^{-\lambda x\delta^2}$$

for all $x > 0$, where $C > 0$, $K > 0$ and $\lambda > 0$ are universal constants.

3. Baseline function estimation in the Cox model

The previous section dealt with a model where the observed sample is (Y_i, δ_i), $i = 1, \dots, n$ with $Y_i = \min(X_i, C_i)$ and $\delta_i = I(X_i \le C_i)$, X_i, C_i are two independent identically distributed (iid.) sequences of random variables. In this section such a model is considered where an additional sequence of positive variables W_i, $i = 1, \dots, n$, is given together with the above pairs (Y_i, δ_i). The triplets (W_i, Y_i, δ_i) are iid., the censoring variable C_i is independent of the pair (W_i, X_i), and we assume the existence of a conditional survival function $S_0^{W_i}(t)$ for fixed X_i and W_i, that is we consider a model, where

$$S_0^{W_i}(t) = P(X_i \ge t \mid W_i) \ ,$$

and $S_0(t)$ is an unknown (deterministic) continuous survival function. We want to estimate this unknown baseline survival function $S_0(t)$ based on a sample of the above triplets. Notice, that in the special case when $W_i = 1$, $i = 1, \dots, n$, this is the censored sample considered in the previous section.

We call this model "nonparametric Cox model", because with appropriate parametrization it provides the semiparametric Cox model. Introducing $W_i = \exp(\beta Z_i)$ where β is an unknown parameter and Z_i is the known regressor variable for $i = 1, \dots, n$ we have the Cox regression model. (See [rt] for more details.)

In paper [rt] the following GMLE type estimator (see previous section) of S_0 was proposed:

$$1 - F_n(t) = \hat{S}_n(t) = \prod_{i=1}^{n} \left(\frac{N(Y_i)}{N(Y_i) + W_i} \right)^{\frac{I(Y_i \le t, \delta_i = 1)}{W_i}} \qquad \text{if } t \le \max(Y_1, \dots, Y_n) \ , \tag{3.1}$$

where

$$N(t) = \sum_{j=1}^{n} W_j \, I(Y_j > t) + \sum_{j=1}^{n} W_j I(Y_j = t, \delta_j = 0) \ .$$

The Kaplan–Meier product limit estimator is a special case of (3.1) when all of the W_i-s equal 1.

The calculations showing that the above nonparametric likelihood estimator is as good as a parametric likelihood estimator is very similar to that given in the previous section. We give an expansion of $\hat{S}_n(t)$ defined in (3.1). The leading term of this expansion is $S_0(t)$, the expression we intend to estimate. The second term is $n^{-1/2}$ times a linear functional of a standardized empirical distribution function, and the remaining error term can be bounded by n^{-1} times a random variable with finite momentum generating function. In such a way, a result can be proved which shows that the estimate (3.1) behaves very similarly to the maximum likelihood estimate in the parametric case. The proof follows the same line as the one in the Kaplan–Meier case.

For the sake of simpler notations we only deal with complete sample, but a similar representation can be given for censored sample. Thus the given sample is (X_i, W_i), $i = 1, \dots, n$, where $P(X_i > t \mid W_i) = S_0^{W_i}(t)$.

We introduce some notations. Put $P(W \le w) = Q(w)$ and

$$\begin{aligned} P(X > t) = \bar{G}(t) = 1 - G(t) &= \int_0^\infty S_0^w(t)\, dQ(w), \\ E\left(W\, I(X > t)\right) = \bar{H}(t) = E(W) - H(t) &= \int_0^\infty w S_0^w(t)\, dQ(w)\ , \\ F(x_1, x_2) &= P(X \le x_1, W \le x_2) \end{aligned} \tag{3.2}$$

Note that

$$F(x_1, x_2) = P(W \le x_2) - P(X > x_1, W \le x_2) = Q(x_2) - \int_0^{x_2} S_0^w(x_1)\, dQ(w)\ , \tag{3.3}$$

and

$$1 - G(t) = \bar{G}(t) = \int_0^\infty \exp\left(w \log(S_0(t))\right)\, dQ(w) = M_W\left(\log(S_0(t)\right), \tag{3.4}$$

where M_W denotes the moment generating function of W. We introduce the empirical processes

$$\begin{aligned} G_n(t) = \frac{1}{n}\sum_{i=1}^n I(X_i \le t), \quad \bar{H}_n(t) &= \frac{1}{n}\sum_{i=1}^n W_i\, I(X_i > t)\ , \\ F_n(x_1, x_2) = \frac{1}{n}\sum_{i=1}^n I(X_i \le x_1, W_i \le x_2). & \end{aligned} \tag{3.5}$$

We suppose that the following conditions hold:

(i) W is a bounded positive random variable with a positive lower bound δ_0, that is $P(K > W > \delta_0) = 1$ with some constant K.

(ii) On the interval $(-\infty, T]$ there exists a fixed positive number κ such that

$$1 - G(t) > \kappa, \quad \forall\, t \in (-\infty, T]\ .$$

We consider $S_0(t) - \hat{S}_n(t)$ on the interval $(-\infty, T]$. It follows from conditions (i) and (ii) that

$$\bar{H}(T) > \delta \quad \text{with some fixed } \delta > 0\ . \tag{3.6}$$

Similarly to the case of the product limit estimator we introduce the cumulative hazard function and its empirical version

$$\Lambda(t) = -\log\left(S_0(t)\right), \quad \Lambda_n(t) = -\log\left(\hat{S}_n(t)\right).$$

Using (3.1) in the case when there is no censoring i.e. when $\delta_i = 1$ for all i, we get that

$$\Lambda_n(t) = -\sum_{i=1}^n \frac{I(X_i \le t)}{W_i} \log\left(1 - \frac{W_i}{N(X_i) + W_i}\right)\ .$$

Using almost the same expansions as in the last section, we also obtain that

$$\sqrt{n}\,(\Lambda_n(t)-\Lambda(t)) = \sqrt{n}B_2(t) - \sqrt{n}B_3(t) - \sqrt{n}B_4(t) + \text{negligible error}, \tag{3.7}$$

where

$$B_2(t) = \frac{1}{\sqrt{n}} \int_{y_1\le t} \int_{\mathbf{R}^2} \frac{x_2\, I(x_1>y_1)}{\bar{H}^2(y_1)}\, dF(x_1,x_2)\, d\mu_n(y_1,y_2)\ ,$$
$$B_3(t) = \frac{1}{\sqrt{n}} \int_{y_1\le t} \int_{\mathbf{R}^2} \frac{x_2\, I(x_1>y_1)}{\bar{H}^2(y_1)}\, d\mu_n(x_1,x_2)\, dF(y_1,y_2)\ ,$$
$$B_4(t) = \frac{1}{n} \int_{y_1\le t} \int_{\mathbf{R}^2} \frac{x_2\, I(x_1>y_1)}{\bar{H}^2(y_1)}\, d\mu_n(x_1,x_2)\, d\mu_n(y_1,y_2)\ ,$$

and where $\mu_n(x_1,x_2) = \sqrt{n}\,(F_n(x_1,x_2)-F(x_1,x_2))$.

This formula is also analogous to (1.6). Again it remains to prove that the term $\sup_{t\le T} B_4(t)$ is also small. Theorem B suggests such an estimate. We have to show that the conditions of Theorem B hold. This time we have the integral of the function $f((x_1,x_2),(y_1,y_2)) = \dfrac{x_2 I(x_1>y_1)}{\bar{H}^2(y_1)}$, and this function equals zero on the diagonal set. It follows from conditions (i)—(ii) and (3.6) that this function is bounded if $y_1 \le T$. This way, Theorem B is applicable. We omit the details, and formulate the final result.

Theorem 3.1 *Let T be such that $\bar{H}(T) > \delta$ with some $\delta > 0$. Then the process $S_0(t) - \hat{S}_n(t)$, $-\infty < t < T$, where $\hat{S}_n(t)$ is defined in formula (3.1) can be represented as*

$$S_0(t) - \hat{S}_n(t) = S_0(t)\,(U(n,t) - V(n,t)) + R(n,t)\ , \quad -\infty < t < T\ ,$$

where

$$\sqrt{n}U(n,t) = \int_{y_1\le t} \frac{d\mu_n(y_1,y_2)}{\bar{H}(y_1)}$$
$$\sqrt{n}V(n,t) = \int_{y_1\le t} \frac{\sqrt{n}(\bar{H}_n(y_1)-\bar{H}(y_1))}{\bar{H}^2(y_1)}\, dF(y_1,y_2)$$

are linear functionals of the empirical processes $\mu_n(y_1,y_2) = \sqrt{n}\,(F_n(y_1,y_2)-F(y_1,y_2))$, $\sqrt{n}(\bar{H}_n(y_1)-\bar{H}(y_1))$, and the error term $R(n,t)$ can be bounded as

$$P\left(\sup_{t\le T} n \mid R(n,t) \mid > x + \frac{C}{\delta}\right) \le Ke^{-\lambda x\delta^2}$$

for all $x>0$, where $C>0$, $K>0$ and $\lambda>0$ are universal constants.

Acknowledgement: The authors would like to thank the referee for his help to improve the presentation of the paper.

REFERENCES

[bhhw] Begun, J.M., Hall, W. J., Huang, W.-M. and Wellner, J. A. (1983) Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics*, **11**, 432–452.

[csr] Csörgő, M. and Révész, P. (1981) *Strong approximations in probability and statistics*, Academic Press, New York, NY.

[gil1] Gill, R. D. (1989) Non- and semi-parametric maximum likelihood estimators and the von Mises method (part I), *Scandinavian Journal of Statistics*, **16**, 97–128.

[gil2] Gill, R. D. (1993) Non- and semi-parametric maximum likelihood estimators and the von Mises method (part II), *Scandinavian Journal of Statistics*, **20**, 271–288.

[joh] Johansen, S. (1978) The product limit estimator as a maximum likelihood estimator, *Scandinavian Journal of Statistics*, **5**, 195–199.

[km] Kaplan, E.L. and Meier P. (1958) Nonparametric estimation from incomplete data, *Journal of American Statistical Association*, **53**, 457–481.

[kw] Kiefer, J. and Wolfowitz, J. (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *Annals of Mathematical Statistics*, **27**, 887–906.

[kmt] Komlós, J., Major, P. and Tusnády, G. (1975) An approximation of partial sums of independent r.v.'s and the sample d.f. I., *Zeitschrift für Wahrscheinlichkeitstheorie und angewandte Gebiete*, **32**, 111–132.

[lc] LeCam, L. (1986) *Asymptotic methods in statistical decision theory*, Springer Verlag.

[mp] Major, P. (1988) On the tail behaviour of the distribution function of multiple stochastic integrals, *Probability Theory and Related Fields*, **78**, 419–435.

[mr] Major, P. and Rejtő, L. (1988) Strong embedding of the distribution function under random censorship, *Annals of Statistics* **16**, 1113–1132.

[mrsi] Major, P. and Rejtő, L. (1997) On the tail behaviour of the distribution function of multiple stochastic integrals in separable metric spaces, (submitted for publication)

[mrt] Major, P., Rejtő, L. and Tusnády, G. (1997) Strong embedding of a baseline function estimator in the Cox model, (manuscript).

[rt] Rejtő, L. and Tusnády, G. (1997) On the Cox regression, *This Volume*.