A consistency result in general censoring models

Sebastian Döhler and Ludger Rüschendorf*

Institute for Mathematical Stochastics, University of Freiburg, Eckerstr. 1, D-79104 Freiburg, Germany

July 4, 2002

Abstract

In this paper we prove a consistency result for sieved maximum likelihood estimators of the density in general random censoring models with covariates. The proof is based on the method of functional estimation. The estimation error is decomposed in a deterministic approximation error and the stochastic estimation error. The main part of the proof is to establish a uniform of large numbers for the conditional log-likelihood functional by using results and techniques from empirical process theory.

Keywords: censoring model, sieved maximum likelihood estimator, functional estimation, empirical process theory

1 Introduction

In this paper we consider a general random censoring model where the survival (failure) time $T \in [0,1]$ is observed exactly only outside a random set Z. In the case of censoring (i.e. when $T \in Z$) the information on T is only that T lies in Z. Special cases are right censored data with Z of the form $[C,1]$, left censored data with $Z = [0,C]$ or interval censored data with $Z = [C_1, C_2]$. We assume that T takes its values in $\mathcal{T} = [0,1]$ and that Z is a measurable random set in $[0,1]$. We also consider a vector of covariables X with values in $\mathcal{X} = [0,1]^k$. The observed data then is given by $(\triangle, Y, X)$ where $\triangle = 1_{\{T \notin Z\}}$ is the indicator of the non-censoring set and $Y = \begin{cases} T \text{ if } \triangle = 1 \\ Z \text{ if } \triangle = 0 \end{cases}$ is the censored variable.

The censored counting process of n iid data then is given by a multiplicative intensity process with intensity of the form $\alpha(t)Y(t)C(t)$ where $C(t)$ denotes the censoring process and the results for the Nelson-Aalen estimator

* Corresponding author. Tel.: +49-761-2035665
E-mail address: ruschen@stochastik.uni-freiburg.de

can be applied in this situation for the estimation of the integrated intensity (see Andersen, Borgan, Gill, and Keiding (1993)).

In this paper we consider the ML-estimator for the density w.r.t. a general class of net sieves. For right censored data consistency results and convergence rates for the ML-estimator were obtained in Kooperberg, Stone, and Truong (1995) based on tensor product splines and in Döhler (1999, 2000) for general net sieves by the method of functional estimation. It turns out that the method of functional estimation can be extended to general censoring models. For some further details we refer to the dissertation of Döhler (2000).

We assume throughout that T and Z are conditionally independent given $X = x$. Further we assume the existence of a continuous conditional density $f_{T|X}$ of T given X bounded away from 0 and ∞ i.e. $0 < c \le f_{T|X} \le C$ for some constants c, C, but we do not pose any smoothness conditions on the underlying density. By our boundedness assumption the underlying density can be parameterized by some bounded function $\alpha(t, x)$,

$$f_{T|X}(t,x) = \frac{\exp\ \alpha(t,x)}{\int_{\mathcal{T}} \exp\ \alpha(u,x)\ du} = f^{\alpha}_{T|X}(t,x) \tag{1.1}$$

Note that this parameterization is not unique. It determines α only up to some additive term depending on x. Therefore, we can expect only a result on consistency for the densities but not for the parametrization itself. The underlying density corresponds to some bounded function α_0 in (1.1). Given a sample $(\delta_1, y_1, x_1), \ldots, (\delta_n, y_n, x_n)$ the conditional log-likelihood is given by

$$\begin{aligned} L_n(\alpha) \quad = \quad & \sum_{i=1}^{n} \Big[\delta_i\ \alpha(y_i, x_i) + (1-\delta_i) \log \int_{z_i} \exp\ \alpha(u, x_i)\ du \qquad (1.2) \\ & \qquad - \log \int_0^1 \exp\ \alpha(u, x_i)\ du \Big]. \end{aligned}$$

where we define $\log \int_{z_i} \exp(\alpha(u, x_i)\ du$ to be zero if $\lambda(z_i) = 0$.

Denote by $\Lambda(\alpha)$ the expected conditional log-likelihood

$$\Lambda(\alpha) = \mathbb{E}\ L_1(\alpha). \tag{1.3}$$

Then $\Lambda(\alpha)$ is maximized at the underlying density $\alpha = \alpha_0$ and

$$|\Lambda(\alpha) - \Lambda(\alpha_0)| = \Lambda(\alpha_0) - \Lambda(\alpha) = \mathbb{E}\ I(\alpha, \alpha_0 | X) \tag{1.4}$$

is the expected Kullback-Leibler distance of the conditional distributions $P_{\alpha|X}$, $P_{\alpha_0|X}$ of $(\triangle, Y)$ given X. So $|\Lambda(\alpha) - \Lambda(\alpha_0)|$ defines a statistical distance between the conditional distributions $P_{\alpha|X}$, $P_{\alpha_0|X}$ integrated by the distribution of X. General convergence results for ML-estimators w.r.t. statistical distances were obtained in several recent papers (see Yang and Barron (1999) and Wong and Shen (1995)). Kooperberg and Clarkson (1997) discuss the empirical and numerical challenges of hazard regression for interval censored data.

The Kullback-Leibler distance is related to the distance in variation by (see e.g. (Rüschendorf 1988, Lemma 4.12))

$$\|P - Q\|^2 \leq 1 - \exp(-I(P|Q)). \tag{1.5}$$

Therefore, convergence of α_n to α_0 in Λ-distance implies stochastic convergence of the total variation distance of the conditional distributions. There are several further relations e.g. to the Hellinger distance (see Wong and Shen (1995, Theorem 5)). We shall prove convergence of the conditional sieved ML-estimator $\widehat{\alpha}_n$ to α_0 w.r.t. the statistical Λ-distance. In the case of right censored models Λ-convergence implies L^1-convergence of the weighted log-hazard functions $\widehat{\lambda}_n$

$$\int_{\mathcal{T}\times\mathcal{X}} \overline{F}_C \; |\widehat{\lambda}_n - \lambda_0| \; dP^{T,X} \to 0 \;\; \text{a.s.} \tag{1.6}$$

where $\widehat{\lambda}_n$, λ_0 are the log-hazard functions corresponding to $\widehat{\alpha}_n$ resp. α_0 and $\overline{F}_C$ is the conditional survival function (Döhler (1999)). In the general case it is not so easy to separate the influence of the censoring from the densities (see Lemma 2.1).

We consider sieves $(\mathcal{F}_n)$ generated by a Vapnik-Cervonenkis class $\mathcal{F}_0 \subset \{f : \mathcal{T} \times \mathcal{X} \to [0,1]\}$ in the following way: Consider for $\beta > 0$, $K \in \mathbb{N}$ the finite sieve class with K nodes and weights bounded by β

$$\begin{aligned} \mathcal{F}(\beta, K) \;\; = \;\; \{\alpha : \mathcal{T} \times \mathcal{X} \to [-K\beta, K\beta], \\ \alpha(t,x) = \sum_{i=1}^{K} c_i f_i(t,x), f_i \in \mathcal{F}_0, |c_i| \leq \beta\}. \end{aligned} \tag{1.7}$$

We assume generally that $\mathcal{F}_0$ is rich enough in the sense that the induced vector space $\mathcal{F} = \langle \mathcal{F}_0 \rangle$ is dense in $(C(\mathcal{T} \times \mathcal{X}), \| \;\|_\infty)$, where $\| \;\|_\infty$ is the sup-norm. For some suitable choice of $\beta_n \uparrow \infty$, $K_n \uparrow \infty$ we will prove Λ-consistency of the sieved ML-estimator

$$\widehat{\alpha}_n = \arg\max_{\alpha \in \mathcal{F}_n} L_n(\alpha) \tag{1.8}$$

over the net $\mathcal{F}_n = \mathcal{F}(\beta_n, K_n)$ i.e. $|\Lambda(\widehat{\alpha}_n) - \Lambda(\alpha_0)| \to 0$. Typical examples of nets generated by VC-classes and satisfying the denseness property are neural nets, wavelet nets, and radial basis function nets.

In section 2 we state the main result of this paper (Theorem 2.2). In section 3 we prove an LLN for the log-likelihood functional (Theorem 3.4). The key part here is a maximal inequality (Proposition 3.1) for which we apply results from empirical process theory. We conclude the paper by mentioning some possible extensions and generalizations of our results. The main tools used in the proofs come from empirical process theory. For easy reference on this topic we have put some of the main notions, definitions and results (covering numbers, VC-class, entropy-integral, etc.) in the appendix. For more details we refer the reader to van der Vaart and Wellner (1996) and to Pollard (1990).

2 Λ-consistency of the sieved ML-estimator

In this section we formulate and prove a Λ-consistency result for the sieved ML-estimator for dense nets generated by VC-classes as in (1.7). In section 3 we separately prove a law of large numbers for the log-likelihood functional which is used in the proof of the consistency result in this section.

The following representation of the Λ-distance will be used for upper bounds.

Lemma 2.1 *For $\alpha \in C\left([0,1] \times [0,1]^k\right)$ holds*

$$\begin{aligned}
&\Lambda(\alpha_0) - \Lambda(\alpha) \\
&\quad = \quad \mathbb{E} I(\alpha, \alpha_0 | X) \\
&\quad = \int_{[0,1]^k} \int_{\mathcal{B}([0,1])} \int_{[0,1]} \Big\{ (\alpha_0(u,x) - \alpha(u,x)) \qquad (2.1) \\
&\qquad + 1_z(u) \left[\alpha(u,x) - \alpha_0(u,x) + \log \int_z \exp \alpha_0(v,x) dv - \log \int_z \exp \alpha(v,x) dv \right] \\
&\qquad - \log \int_0^1 \exp \alpha_0(v,x) dv + \log \int_0^1 \exp \alpha(v,x) dv \Big\} dP^{T,Z,X} \ (u,z,x)
\end{aligned}$$

$$\begin{aligned}
&\quad = \int_{[0,1]^k} \int_{\mathcal{B}([0,1])} \int_{[0,1]} \Big\{ (\alpha_0(u,x) - \alpha(u,x)) \qquad (2.2) \\
&\qquad + 1_z(u) \left[\alpha(u,x) - \alpha_0(u,x) - \log \mathbb{E}_{\mu_{z,x}} \ \exp(\alpha - \alpha_0)\right] \\
&\qquad - \log \int_0^1 \exp \ \alpha_0(v,x) dv + \log \int_0^1 \exp \ \alpha(v,x) dv \Big\} dP^{(T,Z,X)}(u,z,x).
\end{aligned}$$

Here for $z \in \mathcal{B}([0,1])$ and $x \in [0,1]^k$ the expectation $\mathbb{E}_{\mu_{z,x}}$ is calculated w.r.t. the distribution $\mu_{z,x}$ with density $1_z(\cdot) \exp \ \alpha_0(\cdot, x) / \int_{[0,1]} 1_z(u) \exp \ \alpha_0(u,x) \ du$ for z with $\lambda(z) > 0$ while $\log E_{\mu_{z,x}} \exp(\alpha - \alpha_0) = 0$ for $\lambda(z) = 0$.

Proof: For any $\alpha : [0,1] \times [0,1]^k \to \mathbb{R}$ holds

$$\begin{aligned}
&\Lambda(\alpha) \\
&\quad = \mathbb{E} \ L_1(\alpha) \\
&\quad = \mathbb{E} \left[\triangle \alpha(T,X) + (1 - \triangle) \log \int_Z \exp \alpha(u,X) du - \log \int_0^1 \exp \alpha(u,X) du \right] \\
&\quad = \mathbb{E} \left[1_{\{T \notin Z\}} \alpha(T,X) + 1_{\{T \in Z\}} \log \int_Z \exp \alpha(u,X) du - \log \int_0^1 \exp \alpha(u,X) du \right] \\
&\quad = \int_{[0,1]^k} \int_{\mathcal{B}([0,1])} \int_{[0,1]} \Big[(1 - 1_z(t)) \alpha(t,x) + 1_z(t) \log \int_z \exp \alpha(u,x) du \\
&\qquad\qquad - \log \int_0^1 \exp \alpha(u,x) dv \Big] \, dP^{(T,Z,X)} \ (t,z,x) \ ,
\end{aligned}$$

which implies (2.1). For the proof of (2.2) note that for $x \in [0,1]^k$, $z \in \mathcal{B}([0,1])$ with $\lambda(z) > 0$ we have

$$\begin{aligned}
&\log \int_z \exp \alpha_0(v,x)\ dv - \log \int_z \exp \alpha(v,x) dv \\
&= \ -\log \frac{\displaystyle\int_z \exp \alpha(v,x)\ dv}{\displaystyle\int_z \exp \alpha_0(v,x)\ dv} \\
&= \ -\log \int_{[0,1]} \exp\left(\alpha(v,x) - \alpha_0(v,x)\right) \frac{1_z(v) \exp \alpha_0(v,x)}{\displaystyle\int_{[0,1]} 1_z(u) \exp \alpha_0(u,x) du} dv \\
&= \ -\log \mathbb{E}_{\mu_{z,x}} \exp(\alpha - \alpha_0)\ . \qquad \square
\end{aligned}$$

The following theorem on consistency of the sieved ML-estimator in the general censoring model is the main result of this paper.

Theorem 2.2 (Consistency of the sieved ML-estimator.)
Let $\beta_n \uparrow \infty$, $K_n \uparrow \infty$ such that

$$\beta_n K_n^2 \exp(2\beta_n K_n) = O(n^{1/4}). \tag{2.3}$$

Then for $\alpha_0 \in C(\mathcal{T} \times \mathcal{X})$ the net $\mathcal{F}_n = \mathcal{F}(\beta_n, K_n)$ generated by the VC class $\mathcal{F}_0$ increases to the dense subset $\mathcal{F}$ and the sieved ML-estimator $\widehat{\alpha}_n$ is Λ-consistent i.e.

$$|\Lambda(\widehat{\alpha}_n) - \Lambda(\alpha_0)| \to 0 \quad a.s. \tag{2.4}$$

Proof: By a standard decomposition argument (see e.g. Döhler (2000)) we obtain for the ML-estimator

$$\begin{aligned}
|\Lambda(\widehat{\alpha}_n) - \Lambda(\alpha_0)| \ \leq \ & \inf_{\alpha \in \mathcal{F}_n} \Big|\Lambda(\alpha) - \Lambda(\alpha_0)\Big| \\
& + 2 \sup_{\alpha \in \mathcal{F}_n} \left|\frac{1}{n} L_n(\alpha) - \Lambda(\alpha)\right|.
\end{aligned} \tag{2.5}$$

The second stochastic term in the decomposition converges to zero a.s. by the strong law of large numbers in Theorem 3.4 which is proved in section 3.

Since the vectorspace $\mathcal{F}$ generated by $\mathcal{F}_0$, $\mathcal{F} = \langle \mathcal{F}_0 \rangle$ is dense in $C(\mathcal{T} \times \mathcal{X})$ there exists for any $\varepsilon > 0$ some $n \in \mathbb{N}, \alpha_n \in \mathcal{F}_n$ such that $\|\alpha_n - \alpha_0\|_\infty < \varepsilon$. This implies by Lemma 2.1

$$\begin{aligned}
&\Big|\Lambda(\alpha_n) - \Lambda(\alpha_0)\Big| \\
&\leq \ \int_{[0,1]^k} \int_{\mathcal{B}([0,1])} \int_{[0,1]} \bigg(\|\alpha_n - \alpha_0\|_\infty + \Big| \log \mathbb{E}_{\mu_{[0,1],x}} \exp(\alpha_n - \alpha_0) \Big| \\
&\quad + 1_z(u) \Big[\|\alpha_n - \alpha_0\|_\infty + \Big| \log \mathbb{E}_{\mu_{z,x}}\ \exp(\alpha_n - \alpha_0) \Big| \Big] \bigg) dP^{(T,Z,X)}(u,z,x)
\end{aligned}$$

$$
\begin{aligned}
\leq\ & \int_{[0,1]^k}\int_{\mathcal{B}([0,1])}\int_{[0,1]}\Big(\|\alpha_n-\alpha_0\|_\infty+\|\alpha_n-\alpha_0\|_\infty \\
& +1_z(u)\cdot\Big[\|\alpha_n-\alpha_0\|_\infty+\|\alpha_n-\alpha_0\|_\infty\Big]\Big)\ dP^{(T,Z,X)}(u,z,x) \\
\leq\ & 4\varepsilon.
\end{aligned}
$$

This implies the result. □

Theorem 2.2 shows that the function classes considered in this paper can be used to estimate the censored distribution of a survival time under quite general assumptions on the censoring mechanism and on the underlying survival time distribution. In addition it is also indicated how the constraints on the number of basis functions and the magnitude of weights should be chosen in order to achieve consistency. Some generalizations and extensions of this approach to other sieve classes and to linear net classes are further suggested in Remarks 3.3, 3.5.

3 A uniformlaw of large numbers for the log-likelihood functional

In this section we prove an LLN for the log-likelihood functional. The main part of the proof is to establish a maximal inequality for $L_n(\alpha)$ by considering the log-likelihood functional as an empirical process induced by the function class $\mathcal{G}$ defined below to which results from empirical process theory apply. Define for $\beta > 0,\ K \in \mathbb{N}$

$$
\begin{aligned}
&\mathcal{G}(\beta, K) \qquad (3.1)\\
&:= \Big\{g_\alpha : \{0,1\}\times\mathcal{B}([0,1])\times\mathcal{X}\longrightarrow\mathbb{R},\ (\delta,y,x)\longmapsto\delta\alpha(y,x) \\
&\qquad +(1-\delta)\log\int_y\exp\alpha(u,x)\ du-\log\int_0^1\exp\alpha(u,x)du\ \Big|\ \alpha\in\mathcal{F}(\beta,K)\Big\}
\end{aligned}
$$

where again $\log\int_y\exp\alpha(u,x)du$ is defined to be zero if $\lambda(y)=0$. Now we can state the maximal inequality for the log-likelihood functional which is used in the proof of the LLN.

Proposition 3.1 *Let $d=\dim_{VC}(\mathcal{F}_0)$, then there exists a constant $C(d)>0$ such that for all $t>0$*

$$
\begin{aligned}
&P\left(\sup_{\alpha\in\mathcal{F}(\beta,K)}\left|\frac{1}{n}L_n(\alpha)-\Lambda(\alpha)\right|\geq t\right) \qquad (3.2)\\
&= P\left(\sup_{g\in\mathcal{G}(\beta,K)}\left|\frac{1}{n}\sum_{i=1}^n g\left(\delta_i,y_i,x_i\right)-\mathbb{E}\ g(\Delta,Y,X)\right|\geq t\right)\\
&\leq 5\exp\left(-\frac{1}{2}\frac{n}{C(d)\beta K^2(\beta K+2\exp(\beta K))^2}t^2\right).
\end{aligned}
$$

For the proof of (3.2) we use the following generalization of a result of Stone (1991, Lemma 1).

Lemma 3.2 *Let $\alpha, \beta : \mathbb{R} \to [-M, M]$, where $M > 0$, let $z \in \mathcal{B}(\mathbb{R})$ with $0 < \lambda(z) < \infty$ where λ denotes the Lebesgue-measure. Then*

$$\begin{aligned} &\left(\log \int_z \exp \alpha(u)\ d\lambda(u) - \log \int_z \exp \beta(u)\ d\lambda(u)\right)^2 \\ &\quad \le \frac{\exp(4M)}{\lambda(z)} \int_z (\alpha(u) - \beta(u))^2\ d\lambda(u). \end{aligned} \tag{3.3}$$

Proof: Define for $f \in \{\alpha, \beta\}$, $I_f := \int_z \exp f(u)\ d\lambda(u)$. Then by the mean value theorem we obtain

$$(\log I_\alpha - \log I_\beta)^2 = \left[\frac{1}{v} \int_z (\exp \alpha(u) - \exp \beta(u))\ d\lambda(u)\right]^2,$$

where $\exp(-M)\lambda(z) \le \min\{I_\alpha, I_\beta\} \le v \le \max\{I_\alpha, I_\beta\} \le \exp(M)\lambda(z)$. Let Q denote the uniform distribution on z. Then

$$(\log I_\alpha - \log I_\beta)^2 \le \left[\exp(M) \int (\exp \alpha(u) - \exp \beta(u))\ dQ(u)\right]^2.$$

From the mean value theorem we obtain

$$(\log I_\alpha - \log I_\beta)^2 \le \left[\exp(M) \int (\exp g(u)) \cdot (\alpha(u) - \beta(u))\ dQ(u)\right]^2$$

where $-M \le \min\{\alpha(u), \beta(u)\} \le g(u) \le \max\{\alpha(u), \beta(u)\} \le M$. Then the Cauchy-Schwarz inequality implies

$$\begin{aligned} &(\log I_\alpha - \log I_\beta)^2 \\ &\quad \le \exp(2M) \left[\int \exp(2g(u))\ dQ(u)\right] \left[\int (\alpha(u) - \beta(u))^2\ dQ(u)\right], \end{aligned}$$

which yields (3.3) as $g \le M$. □

Proof of Proposition 3.1: The idea of the proof is to show that the random entropy integral of $\mathcal{G}(\beta, K)(cf.\ (3.12))$ can be bounded by a deterministic function of β and K and to subsequently use Pollard's maximal inequality for means of random processes (cf. (4.6)). Our first goal is to obtain a bound on the covering numbers for the log-likelihood functional.

For $\beta > 0, K \in \mathbb{N}$ define using our usual convention on integrals over sets y of measure zero

$$\begin{aligned} \mathcal{H}(\beta, K) &:= \{h_\alpha : \{0,1\} \times \mathcal{B}([0,1]) \times \mathcal{X} \to \mathbb{R}, \\ &\qquad (\delta, y, x) \mapsto \delta\alpha(y, x) \big| \alpha \in \mathcal{F}(\beta, K)\} \\ \mathcal{K}(\beta, K) &:= \{k_\alpha : \{0,1\} \times \mathcal{B}([0,1]) \times \mathcal{X} \to \mathbb{R}, \\ &\qquad (\delta, y, x) \mapsto (1-\delta) \log \int_y \exp \alpha(u, x)\ du \big| \alpha \in \mathcal{F}(\beta, K)\} \\ \mathcal{L}(\beta, K) &:= \{\ell_\alpha : \{0,1\} \times \mathcal{B}([0,1]) \times \mathcal{X} \to \mathbb{R}, \\ &\qquad (\delta, y, x) \mapsto -\log \int_0^1 \exp \alpha(u, x)\ du \big| \alpha \in \mathcal{F}(\beta, K)\}, \end{aligned}$$

then $\mathcal{G}(\beta, K) \subset \mathcal{H}(\beta, K) \bigoplus \mathcal{K}(\beta, K) \bigoplus \mathcal{L}(\beta, K)$. Note that formally we should define the variable y in $(\mathcal{B}[0,1] \cup [0,1])$ corresponding to the censored resp. uncensored case but we identify $\{y\}$ with the variable y in the following. Again we can restrict to the case of $y_i \in \mathcal{B}([0,1])$ with $\lambda(y_i) > 0$. Let ν_n denote the empirical measure generated by the sample $\{(\delta_1, y_1, x_1), \cdots, (\delta_n, y_n, x_n)\}$, $\nu_n := \frac{1}{n}\sum_{i=1}^n \delta_{(\delta_i, y_i, x_i)}$, and define $\widetilde{\nu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{(y_i, x_i)}$, $\widetilde{\widetilde{\nu}}_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Then from some composition rules for covering numbers (see Lemma 4.4) we obtain

$$
\begin{aligned}
&N\left(\varepsilon, \mathcal{G}(\beta, K), d_{L^2(\nu_n)}\right) \qquad (3.4)\\
&\quad \leq N\left(\frac{\varepsilon}{3}, \mathcal{H}(\beta, K), d_{L^2(\nu_n)}\right) N\left(\frac{\varepsilon}{3}, \mathcal{K}(\beta, K), d_{L^2(\nu_n)}\right) N\left(\frac{\varepsilon}{3}, \mathcal{L}(\beta, K), d_{L^2(\nu_n)}\right).
\end{aligned}
$$

Obviously,

$$
N\left(\varepsilon, \mathcal{H}(\beta, K), d_{L^2(\nu_n)}\right) \leq N\left(\varepsilon, \mathcal{F}(\beta, K), d_{L^2(\widetilde{\nu}_n)}\right). \qquad (3.5)
$$

Let R_i denote the uniform distribution on z_i, if $\lambda(z_i) > 0$ and define $Q_n := \frac{1}{n}\sum\limits_{i=1}^n R_i \bigotimes \delta_{x_i}$. Then

$$
N\left(\varepsilon, \mathcal{K}(\beta, K), d_{L^2(\nu_n)}\right) \leq N\left(\frac{\varepsilon}{\exp 2M}, \mathcal{F}(\beta, K), d_{L^2(Q_n)}\right). \qquad (3.6)
$$

For the proof of (3.6) observe that for $k_{\alpha_1}, k_{\alpha_2} \in \mathcal{K}(\beta, K)$ using Lemma 3.2

$$
\begin{aligned}
d^2_{L^2(\nu_n)}(k_{\alpha_1}, k_{\alpha_2}) &\leq \frac{1}{n}\sum_{i=1}^n \left(\log \int_{y_i} \exp \alpha_1(u, x_i) du - \log \int_{y_i} \exp \alpha_2(u, x_i)\, du\right)^2 \\
&\leq \exp(4M) \cdot \frac{1}{n}\sum_{i=1}^n \frac{1}{\lambda(y_i)} \int_{y_i} (\alpha_1(u, x_i) - \alpha_2(u, x_i))^2 \lambda(du) \\
&= \exp(4M) \cdot \int_{[0,1]\times[0,1]^k} (\alpha_1 - \alpha_2)^2 dQ_n.
\end{aligned}
$$

Similarly with $Z = [0,1]$ we obtain

$$
N\left(\varepsilon, \mathcal{L}(\beta, K), d_{L^2(\nu_n)}\right) \leq N\left(\frac{\varepsilon}{\exp 2M}, \mathcal{F}(\beta, K), d_{L^2(\widetilde{\widetilde{\nu}}_n \bigotimes U[0,1])}\right). \qquad (3.7)
$$

(3.4) – (3.7) imply

$$
\begin{aligned}
&N(\varepsilon, \mathcal{G}(\beta, K), d_{L^2(\nu_n)}) \qquad (3.8)\\
&\quad \leq N\Big(\frac{\varepsilon}{3}, \mathcal{F}(\beta, K), d_{L^2(\widetilde{\nu}_n)}\Big) N\Big(\frac{\varepsilon}{3\exp 2M}, \mathcal{F}(\beta, K), d_{L^2(Q_n)}\Big) \\
&\qquad \cdot N\Big(\frac{\varepsilon}{3\exp 2M}, \mathcal{F}(\beta, K), d_{L^2(\widetilde{\widetilde{\nu}}_n \otimes U[0,1])}\Big).
\end{aligned}
$$

From Pollard's estimate for the covering number of majorized VC-classes (see (4.1)), we obtain for any probability measure ν on $\mathcal{T} \times \mathcal{X}$ and any $\delta > 0$

$$
N(\delta, \mathcal{F}(\beta, K), d_{L^2(\nu)}) \leq C(d)^K (\beta K)^{K(2d-1)} \left(\frac{1}{\delta}\right)^{K(2d-1)}. \qquad (3.9)
$$

This estimate will be used in particular for the measures $\widetilde{\nu}_n$, Q_n, and $\widetilde{\widetilde{\nu}}_n \otimes \mathcal{U}[0,1]$ above. To apply Pollard's maximal inequality we have to show that the random entropy integral of $\mathcal{G}(\beta, K)$ is bounded by a constant (cf. (4.6)). To that purpose we estimate the L_2-packing numbers of the random point set

$$\mathcal{Z}_n(\mathcal{G}(\beta, K))(\omega) = \{(\Phi_1(\alpha, \omega), \dots, \Phi_n(\alpha, \omega) \mid \alpha \in \mathcal{G}(\beta, K)\} \tag{3.10}$$

where $\Phi_i(\alpha, \omega) = \Phi_i(g_\alpha, \omega) = g_\alpha(\delta_i, y_i, x_i)$. For the definition of covering resp. packing numbers N_2, D_2 see Pollard (1990, section 4)

$$\begin{aligned} D_2(\varepsilon, \mathcal{Z}_n(\mathcal{G}(\beta, K))(\omega)) &\leq N_2\left(\frac{\varepsilon}{2}, \mathcal{Z}_n(\mathcal{G}(\beta, K))(\omega)\right) \\ &= N\left(\frac{\varepsilon}{2}, \mathcal{G}(\beta, K), \sqrt{n} d_{L^2(\nu_n)}\right) \\ &\leq N\left(\frac{\varepsilon}{2\sqrt{n}}, \mathcal{G}(\beta, K), d_{L^2(\nu_n)}\right). \end{aligned}$$

Using stability properties of covering numbers (as in Lemma 4.4) we obtain

$$\begin{aligned} N(\delta, \mathcal{G}(\beta, K), d_{L^2(\nu_n)}) &\leq N(\delta, \mathcal{F}(\beta, K), d_{L^2(\widetilde{\nu}_n)}) \\ &\leq N\left(\frac{\delta}{K} \bigoplus_{i=1}^{K} [-\beta, \beta] \cdot \mathcal{F}_0, d_{L^2(\widetilde{\nu}_n)}\right) \\ &\leq \left(N\left(\frac{\delta}{K}[-\beta, \beta] \cdot \mathcal{F}_0, d_{L^2(\widetilde{\nu}_n)}\right)\right)^K \\ &\leq \left(N\left(\frac{\delta}{2\beta K}, \mathcal{F}_0, d_{L^2(\widetilde{\nu}_n)}\right) \frac{4\beta K}{\delta}\right)^K \\ &\leq \left(C(d)\left(\frac{\delta}{2\beta K}\right)^{-2(d-1)} \frac{4\beta K}{\delta}\right)^K \\ &= \widetilde{C}(d)(\beta K)^{K(2d-1)}\left(\frac{1}{\delta}\right)^{K(2d-1)}. \end{aligned}$$

For the last inequality we use Pollard's estimate of the covering numbers of a VC-class (see (4.1)). As consequence we obtain

$$\begin{aligned} &D_2(\varepsilon, \mathcal{Z}_n(\mathcal{G}(\beta, K))(\omega)) \\ &\leq C(d)^K (\beta K)^{3K(2d-1)}\Big(\exp(4\beta K)\Big)^{K(2d-1)}\left(\frac{\sqrt{n}}{\varepsilon}\right)^{3K(2d-1)}. \end{aligned} \tag{3.11}$$

With $a_n = \sqrt{n}(\beta K + 2\exp \beta K)$ and with the substitution $y = \frac{x}{a_n}$ this implies for the random entropy integral

$$\begin{aligned} \mathcal{J}_n(\mathcal{G}(\beta, K))(\omega) &= 9\int_0^{\delta_n(\omega)} \sqrt{\log D_2(x, \mathcal{Z}_n(\mathcal{G}(\beta, K))(\omega))}\, dx, \qquad (3.12) \\ &\leq 9a_n \int_0^1 \sqrt{\log D_2(a_n y, \mathcal{Z}_n(\mathcal{G}(\beta, K))(\omega))}\, dy \end{aligned}$$

here $\delta_n(\omega) = \sup\{\|y\|_2 \,|\, y \in \mathcal{Z}_n(\mathcal{G}(\beta, K))\}$. Therefore, the random entropy integral is bounded by the deterministic function $C(d)\sqrt{n}K\sqrt{\beta}(\beta K + 2\exp(\beta K))$.

Now we apply Pollard's maximal inequality (see (4.6)) to obtain

$$
\begin{aligned}
& P\Big(\sup_{g\in\mathcal{G}(\beta,K)} \Big|\frac{1}{n}\sum_{i=1}^{n} g(\delta_i, Z_i, y_i, x_i) - \mathbb{E}\Big[g(\delta, Z, Y, X)\Big]\Big| \geq t\Big) && (3.13)\\
\leq\; & 5\exp\Big(-\frac{1}{2}\frac{n}{(C(d)\sqrt{\beta}K(\beta K + 2\exp\beta K))^2}\, t^2\Big)\\
=\; & 5\exp\Big(-\frac{1}{2}\frac{n}{C(d)\beta K^2(\beta K + 2\exp(\beta K))^2}\, t^2\Big).
\end{aligned}
$$

□

Remark 3.3 *Note that the proof of Proposition 3.1 does not need exactly that $\mathcal{F}$ is generated by a VC-class.* (3.8) *holds true for any uniformly bounded class $\mathcal{F}$ with $\mathcal{G} = \{g_f \,|\, f \in \mathcal{F}\}$, g_f as in* (3.1)*. For the bound on the random entropy integral we need that for any probability measure ν the covering numbers $N(\varepsilon, \mathcal{F}, d_{L^2(\nu)})$ grow polynomially in $\frac{1}{\varepsilon}$ (as is true in particular for VC-classes).*

As consequence we now obtain

Theorem 3.4 (Strong law for the likelihood functional.)
Let $\beta_n \uparrow \infty$, $K_n \uparrow \infty$ solve $\beta_n K_n^2 \exp(2\beta_n K_n) = O\left(n^{\frac{1}{4}}\right)$, then the strong law holds for the log-likelihood functional and the sieve $\mathcal{F}_n = \mathcal{F}(\beta_n, K_n)$

$$
\sup_{\alpha\in\mathcal{F}_n}\Big|\frac{1}{n}L_n(\alpha) - \Lambda(\alpha)\Big| \to 0 \quad a.s. \tag{3.14}
$$

Proof: Define

$$
A_n := \left\{\sup_{\alpha\in\mathcal{F}_n}\Big|\frac{1}{n}L_n(\alpha) - \Lambda(\alpha)\Big| \geq \frac{1}{n^{\frac{1}{4}}}\right\}.
$$

Then from Proposition 3.1 we obtain for n sufficiently large $P(A_n) \leq 5\exp\left(-Cn^{\frac{1}{4}}\right)$. Since $\exp\left(-Cn^{\frac{1}{4}}\right) \leq \frac{1}{n^2}$ for large n we obtain $\sum_{n\in\mathbb{N}} P(A_n) < \infty$ and the result follows from the Borel-Cantelli Lemma. □

Remark 3.5 *a) The consistency result in this paper which is stated for rich nets generated by VC-classes can be proved in a similar way also for linear nets given by a sequence $(V_n) \subset \{f : \mathcal{T}\times\mathcal{X} \to \mathbb{R}\}$ of finite-dimensional linear subspaces with the denseness condition that any continuous function f can be approximated uniformly by a sequence $f_n \in V_n$. Examples for* *linear sieve classes are polynomial classes (with V_N generated by N basis functions) and tensor product splines. Defining the truncation operator*

$$T_\beta(x) = \begin{cases} \beta & for \quad x > \beta \\ x & for \ -\beta \le x \le \beta \\ -\beta & for \ -\beta > x \end{cases}$$

then one can prove similarly to the case of nets Λ-convergence of the sieved ML-estimator w.r.t. the sieve

$$\mathcal{F}_n = \Phi(\beta_n, V_{N_n}) = T_{\beta_n} \circ V_{N_n} \tag{3.15}$$

where $\beta_n \uparrow \infty$, $K_n = \dim V_{N_n} \uparrow \infty$ ($\dim V_{N_n}$ the algebraic dimension of V_{N_n}) are growing slowly. For more details in the case of right-censoring see Döhler (1999, 2000).

b) *Under smoothness assumptions on the underlying conditional log-hazard function Kooperberg, Stone, and Truong (1995) established a convergence rate for the tensorproduct spline estimator in the right censored case. For general sieve type estimators related convergence rate results under various 'complexity' conditions on the model are given in Döhler and Rüschendorf (2000). It seems possible to extend these results to the present general censoring case. In this paper the main focus is on a universal consistency result which needs only weak regularity assumptions on the underlying density.*

4 Appendix: Some results from empirical process theory

Here we introduce some results and notation we need to derive probability bounds on the maximal deviation of empirical processes indexed by functions. For the first two definitions cf. van der Vaart and Wellner (1996).

Definition 4.1 (Covering Numbers, Packing Numbers)
Let (T,d) be a semimetric space. Then the covering number *$N(\varepsilon, T, d)$ is defined as the minimal number of balls of radius ε needed to cover T. The* packing number *$D(\varepsilon, T, d)$ is defined as the maximal number of points in T such that the distance between each pair of points is strictly larger than ε.*

Definition 4.2 (Vapnik-Cervonenkis-class)
Let $\mathcal{C}$ be a collection of subsets of some set $\mathcal{S}$. A set $s \subset \mathcal{S}$ with $s = \{s_1, \dots, s_n\}$ is shattered *by $\mathcal{C}$ if for any $s' \subset s$ there is a $c' \in \mathcal{C}$ such that*

$s' = s \cap c'$. *The* VC-index *of* $\mathcal{C}$ *is the smallest* $n \in \mathbb{N}$ *such that for no set* $s \in \mathcal{S}$ *with* $|s| = n$ s *can be shattered by* $\mathcal{C}$.

A collection $\mathcal{F}$ *of measurable functions on some space* $\mathcal{Y}$ *is called a VC-class of functions (and the associated VC-index is called the VC-dimension) if the collection of all subgraphs of the functions (that is sets* $\{(x,t) : t < f(x)\}$ *for* $f \in \mathcal{F}$*) in* $\mathcal{F}$ *is a VC-class of sets in* $\mathcal{Y} \times \mathbb{R}$.

Lemma 4.3 (Pollard)
Let $\mathcal{F}$ *be a Vapnik-Cervonenkis-class with envelope function* F *and* $d = \dim_{VC}(\mathcal{F})$. *Then there is a constant* $C(d) \geq 0$ *such that*

$$N(\varepsilon \|\mathrm{F}\|_{L^2(\mu)}, \mathcal{F}, d_{L^2(\mu)}) \leq C(d)\varepsilon^{-2(d-1)} \tag{4.1}$$

for all $0 < \varepsilon \leq 1$ *and all probability measures* μ.

Proof: cf. Theorem 2.6.7. in van der Vaart and Wellner (1996). □

Lemma 4.4 (Stability properties of covering numbers)
Let μ *be a probability measure and let* $\mathcal{F}, \mathcal{G}$ *function classes on* $\mathbb{R}^m$. *Then*

a.
$$N(\varepsilon + \delta, \mathcal{F} \oplus \mathcal{G}, d_{L^2(\mu)}) \leq N(\varepsilon, \mathcal{F}, d_{L^2(\mu)}) \cdot N(\delta, \mathcal{G}, d_{L^2(\mu)}) \tag{4.2}$$

b. If $\forall f \in \mathcal{F} : |f| \leq K$ *then for* $a > 0$

$$N(\varepsilon, [-a,a] \cdot \mathcal{F}, d_{L^2(\mu)}) \leq N\left(\frac{\varepsilon}{2a}, \mathcal{F}, d_{L^2(\mu)}\right) \cdot \frac{4aK}{\varepsilon} \tag{4.3}$$

c. Let $\forall f \in \mathcal{F} : |f| \leq K$ *and* $\varphi : [-K, K] \longrightarrow \mathbb{R}$ *be a Lipschitz-function with Lipschitz-constant Lip*(φ)*, then:*

$$N(\varepsilon, \varphi \circ \mathcal{F}, d_{L^2(\mu)}) \leq N\left(\frac{\varepsilon}{\mathrm{Lip}(\varphi)}, \mathcal{F}, d_{L^2(\mu)}\right) \tag{4.4}$$

d. If $\forall f \in \mathcal{F} : |f| \leq K$ *and* $\forall g \in \mathcal{G} : |g| \leq L$*, then:*

$$N(\varepsilon, \mathcal{F} \odot \mathcal{G}, d_{L^2(\mu)}) \leq N\left(\frac{\varepsilon}{2L}, \mathcal{F}, d_{L^2(\mu)}\right) \cdot N\left(\frac{\varepsilon}{2K}, \mathcal{G}, d_{L^2(\mu)}\right) \tag{4.5}$$

Proof: Similar to the calculations in Section 5 in Pollard (1990) and Chapter 29 in Devroye, Györfi, and Lugosi (1996). □

Next we introduce some terminology from Section 7 in Pollard (1990).

Definition 4.5 *For a class of functions $\mathcal{F}$ let $\Phi_i : \mathcal{F} \times \Omega \longrightarrow \mathbb{R}$, $i = 1, ..., n$ be independent stochastic processes. Define the*

Random set of points indexed by $\mathcal{F}$

$$\mathcal{Z}_n(\mathcal{F})(\omega) \;:=\; \Big\{\Big(\Phi_1(\alpha,\omega), \dots, \Phi_n(\alpha,\omega)\Big)\Big|\; \alpha \in \mathcal{F}\Big\} \subset \mathbb{R}^n$$

and the

Random entropy integral

$$\begin{aligned} J_n(\mathcal{F})(\omega) &:= 9\int_0^{\delta_n(\omega)} \sqrt{\log D_2(x, \mathcal{Z}_n(\mathcal{F})(\omega))}\; dx \\ \text{where} \qquad \delta_n(\omega) &:= \sup\{\ \|y\|_2 \mid y \in \mathcal{Z}_n(\mathcal{F})(\omega)\}. \end{aligned}$$

Proposition 4.6 (Pollards maximal inequality)
For stochastic processes Φ_i as in Definition 4.5 assume that the random entropy integral $J_n(\mathcal{F})$ is bounded by a constant K_n, then for all $t > 0$ holds

$$P\left(\sup_{t\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}[\Phi_i(f,\omega) - E\Phi_i(f,\omega)\right| \geq t\right) \leq 5\exp\left(-\frac{n^2t^2}{2K_n^2}\right) \tag{4.6}$$

cf. Pollard (1990, (7.3)).

References

Andersen, P. K., O. Borgan, R. Gill, and N. Keiding (1993). *Statistical models based on counting processes.* Springer.

Devroye, L., L. Györfi, and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, New York.

Döhler, S. (1999). Consistent hazard regression estimation by sieved maximum likelihood estimators. In *Proceedings Conference on Limit Theorems Balatonlelle.* to appear.

Döhler, S. (2000). *Empirische Risiko-Minimierung bei zensierten Daten.* Dissertation, Universität Freiburg.

Döhler, S. and L. Rüschendorf (2000). Adaptive estimation of hazard functions. Preprint, Universität Freiburg.

Kooperberg, C. and D. B. Clarkson (1997). Hazard regression with interval-censored data. *Biometrics 53*, 1485–1494.

Kooperberg, C., C. J. Stone, and Y. K. Truong (1995). The L_2 rate of convergence for hazard regression. *Scandinavian Journal of Statistics 22*, 143–157.

Pollard, D. (1990). *Empirical Processes: Theory and Applications.* Institute of Mathematical Statistics, Hayward.

Rüschendorf, L. (1988). *Asymptotische Statistik.* Teubner Skripten zur Mathematischen Stochastik.

Stone, C. J. (1991). Asymptotics for doubly flexible logspline response models. *Annals Statistics 19*, 1832–1854.

van der Vaart, A. and J. A. Wellner (1996). *Weak convergence and empirical processes.* Springer, New York.

Wong, W. H. and X. Shen (1995). Probability inequalities for Likelihood ratios and convergence rates of sieve MLES. *Annals Statistics 23*, 339–362.

Yang, Y. and A. R. Barron (1999). Information-theoretic determination of minimax rates of convergence. *Annals Statistics 27*, 1564–1599.